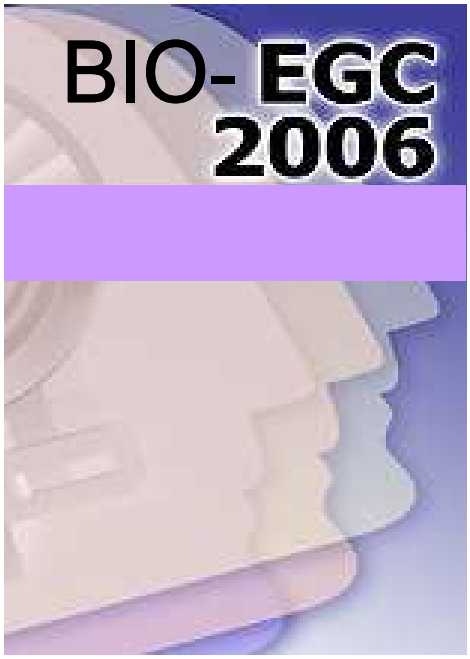

Atelier EGC 2006
17 janvier 2006
Lille
ENIC, Villeneuve d'Ascq

Extraction et Gestion des Connaissances appliquées aux données biologiques

PRHoD : un outil de classification de protéines pour la recherche d'homologues distants

Nicolas Beaume[*,**], Gérard Ramstein[**]
Jérôme Mikolajczak[*,**], Yannick Jacques[*]

[*]INSERM U601 Département de Cancérologie, institut de biologie,
équipe cytokines et récépteurs
9 quai Moncousu F-44035 Nantes cedex
nicolas.beaume@univ-nantes
[**]LINA équipe C.O.D, École polytechnique de l'université de Nantes
rue Christian Pauc
gerard.ramstein@univ-nantes.fr

Résumé. Cet article décrit une méthode d'extraction de nouvelles protéines. Cette méthode utilise cinq classifieurs, dont un original, adaptés aux séquences biologiques et appliquant la technique des Séparateurs à Vastes Marges (SVM). Ces classifieurs se basent notamment sur la composition des séquences en mots de taille fixe, sur la présence de motifs extraits d'un jeu de référence ou encore sur des scores de similarité. Cette méthode a été implémentée dans PRHoD, un outil de classification utilisable sur n'importe quel ensemble de séquences. Elle a été confronté à des données bruitées pour évaluer sa robustesse.

1 Introduction

Certains problèmes biologiques, comme la recherche de séquences homologues, peuvent être assimilés à des problèmes de classification. Deux séquences sont dites homologues si elles proviennent d'une même séquence ancestrale. L'homologie est une information importante pour la connaissance des séquences et extraire cette connaissance des banques de données requiert des techniques de plus en plus rapides et efficaces. Comme il est difficile de déterminer directement l'homologie à cause de la différenciation des séquences au cours du temps, celle-ci est inférée indirectement. Les stratégies les plus courantes sont basées sur la similarité de séquences [Altschul et al. (1997)], les chaînes de Markov cachées [Karplus et al. (1998)] ou les Séparateurs à Vastes Marges (SVM) [Leslie et al. (2002)]. La principale difficulté est de retrouver les homologues quand les séquences sont peu similaires, comme c'est le cas dans la famille des cytokines. Pour des familles particulièrement hétérogènes comme celle-ci, les SVM se sont avérés plus efficaces que les autres méthodes [Leslie et al. (2002)].
Cet article présente une stratégie de recherche de nouveaux membres dans des familles de gènes. Cette stratégie utilise une combinaison de cinq classifieurs basés sur les SVM. Elle est implémentée dans PRHoD, un logiciel supervisant l'ensemble des opérations de recherche de nouveaux membres de familles de gènes. Nous avons testé les performances des classifieurs et démontrons, entre autres, qu'ils sont robustes vis à vis de données bruitées.

2 Classifieurs

Les SVM [Vapnik (1998)] sont une technique d'apprentissage supervisé qui a fait ses preuves dans plusieurs domaines, y compris la bioinformatique. Le principe de base de cette technique est de transformer les données en vecteurs (phase dite de "vectorisation") et de projeter ces vecteurs dans un espace de grande dimension pour déterminer l'hyperplan de séparation entre les membres de la classe d'intérêt et les autres données. La grande force de cette théorie est de proposer de maximiser la distance entre l'hyperplan et les exemples/contre-exemples les plus proches (vecteurs supports). Plusieurs méthodes de vectorisation ont été proposées pour des séquences biologiques. Nous présentons quatre d'entre elles ainsi qu'une cinquième, baptisée HMotifs, que nous avons développée.

2.1 HMotifs

Ce classifieur a été mis au point par Mikolajczak et al. (2004). Il utilise des motifs spécifiques pour vectoriser les séquences, ces motifs étant obtenus lors de l'apprentissage.
Soit Ω l'ensemble des acides aminés représenté par un alphabet de 20 lettres. On peut élargir cet alphabet aux classes de Taylor (1986) qui regroupent les acides aminés selon leurs propriétés physico-chimiques (un même acide aminé pouvant appartenir à plusieurs classes). Par exemple il existe une classe qui représente l'ensemble des aliphatiques {I, L, V}. En y ajoutant Ω lui même, on obtient un nouvel alphabet E(Ω) :

$$E(\Omega) = \{\{A\}, \{C\}, \{D\}, \ldots, \{I, L, V\}, \{H, K, R\}, \{D, E, H, K, R\}, \ldots, \Omega\}$$

On considérera l'ensemble ordonné (E(Ω), $\subseteq$) qui forme un sup-demi-treillis et on notera *sup*(x,y) la borne supérieure de la paire (x, y) de E(Ω) $\times$ E(Ω). Un motif m est un k-mot constitué d'ensembles de E(Ω). Le support d'un motif est le nombre de séquences qui possédent le motif dans le jeu de séquence. La spécificité d'un motif est mesurée à l'aide de la fonction de coût :

$$c(m) = \prod_{j=1}^{k} f(m_j)$$

où $f(m_j)$ est la fréquence de la classe physico-chimique m_j dans l'ensemble des contre-exemples (comprenant 6615 séquences). La spécificité d'un motif m est définie par :

$$\phi(m) = -log(c(m))$$

Soit deux motifs m^1 et m^2, on notera : $m^1 \preceq m^2$ ssi $\forall i \in [1, k]$ on a $m_i^1 \subseteq m_i^2$. La spécificité $\phi(m)$ est telle que $\phi(m^1) \geq \phi(m^2)$ pour toutes les paires vérifiant $m^1 \preceq m^2$. La borne supérieure de m^1 et m^2 est le motif $m^{1,2}$ vérifiant $m^{1,2} = sup(m_i^1, m_i^2) \forall i \in [1, k]$. Le motif $m^{1,2}$ est une généralisation des motifs m^1 et m^2 : tout k-mot vérifiant m^1 ou m^2 vérifiera $m^{1,2}$.
L'algorithme est initialisé avec des motifs germes constitués uniquement de singletons, puis les motifs généralisés sont selectionnés comme il est indiqué ci-dessous :

Entrées :

M : ensemble des motifs germes
supMin : seuil de support minimal toléré
speMin : seuil de spécificité minimal toléré
Sortie :
E : ensemble des motifs généralisés

$E = \{m \in M, support(m) \geq supMin\, et\, \phi(m) \geq speMin\}$
répéter
soit m^1 et m^2 la paire de motifs telle que :
$m^{1,2}$ = sup(m^1, m^2) et $\phi(m^{1,2}) \geq \phi(m^{i,j})\, \forall m^i\, et\, m^j \in M$
$M \leftarrow M - \{m^1, m^2\}$
$M \leftarrow M \cup \{m^{1,2}\}$
si $support(m^{1,2}) \geq supMin\, et\, \phi(m^{1,2}) \geq speMin$
alors $E \leftarrow E \cup \{m^{1,2}\}$
Jusqu'à $cardinal(M) = 1\, ou\, \phi(m^{1,2}) < speMin$

Environ 700 motifs de taille 8 ont été retenus de cette façon pour les cytokines en utilisant un support minimal de 2 et un seuil de spécificité de 14. Les séquences sont représentées sous la forme de vecteurs indiquant la présence du motif de rang i, $i = 1 \ldots |E|$ dans la séquence considéré.

2.2 Autres classifieurs

Spectrum
Cette méthode développée par Leslie et al. (2002) est basée sur la composition en k-mots de la séquence. Son principal avantage est sa rapidité d'exécution.
Mismatch
Ce classifieur [Leslie et al. (2004)] est un raffinement de Spectrum. Il s'appuie toujours sur la composition en k-mots mais tolère un certain nombre de mésappariements. Ce classifieur est moins rapide que Spectrum mais il est connu pour donner de meilleurs résultats.
Pairwise
Ce classifieur [Liao et Noble (2003)] vectorise à partir de scores de similarités. La séquence à vectoriser est alignée contre les séquences de la famille d'interêt et les scores SW (score de Smith & Watermann) forment les composantes du vecteur. Pairwise est plus lent mais plus performant que Spectrum et Mismatch.
LAkernel
LA kernel est également un classifieur basé sur le calcul d'un score de similarité mais qui vérifie toutes les propriétés mathématiques d'une fonction noyau. Saigo et al. (2004) ont démontré que le score SW n'est pas une fonction noyau théoriquement valide pour les SVM. LA kernel a été développé pour respecter ces propriétés tout en conservant la performance de Pairwise. D'un point de vue pratique, LA kernel peut être implémenté comme méthode de vectorisation ou comme une fonction noyau. Saigo *et al* ont montré que ces deux techniques conduisent à des résultats équivalents. Pour garder un traitement des données cohérent avec les autres classifieurs, nous avons implémenté LA kernel en tant que méthode de vectorisation. Ce classifieur est plus lent mais bien plus efficace que Spectrum ou Mismatch.

3 PRHoD (Plateforme de Recherche d'Homologues Distants)

Notre stratégie de recherche est implémentée dans une plateforme qui gére la plupart des aspects de la détection d'homologues : récupération des données, stockage des données, création des entrées nécessaires aux classifieurs, classification et stockage des résultats. PRHoD se présente sous la forme d'un programme java et d'une base de données postgreSQL. La figure

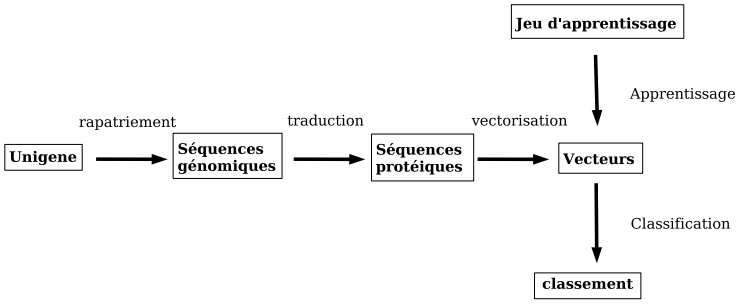

FIG. 1 – *flux de données dans PRHoD*

1 montre comment l'enchainement des traitements dans PRHoD : Des séquences nucléiques humaines sont rapatriées depuis la banque de données Unigene [1] qui contient environ cinq millions de séquences humaines. Les séquences sont traduites en séquences protéiques qui sont utilisées comme entrées pour les cinq classifieurs de la plateforme. Chaque classifieur rend en sortie un classement des séquences de la séquence la plus proche de la famille d'intéret à la plus éloignée. Le classement des séquences est basé sur leur distance à la marge, qui peut être considérée comme un indicateur de confiance de la classification.
L'algorithme SVM utilisé par la plateforme est libsvm [2].

4 Application

4.1 Apprentissage et validation croisée

L'apprentissage a été effectué avec un jeu de 45 séquences de cytokines et 45 contre-exemples tirés de la base SCOP. Les résultats de validation croisée à 10% sont présentés dans la table 1.

4.2 Test sur des séquences génomiques

En vue de tester la robustesse des classifieurs, nous avons classé des séquences bruitées. Pour obtenir ces séquences bruitées, nous avons récupéré les séquences de cytokines d'Unigene et conservé celles qui étaient le plus similaires aux cytokines du jeu d'apprentissage, soit

[1] http : //www.ncbi.nlm.nih.gov/entrez/query.fcgi ?db=unigene
[2] disponible sur le site : http ://www.csie.ntu.edu.tw/c̆jlin/libsvm/

classifieur	vrai positifs	faux négatifs	vrai nég.	faux pos.	% bien classé
spectrum	79%	21%	84%	16%	**82%**
mismatch	87 %	13 %	87 %	13 %	**87%**
HMotifs	100%	0 %	100 %	0%	**100%**
pairwise	100%	0 %	98 %	2 %	**99%**
LAkernel	98%	2%	100 %	0%	**99%**

TAB. 1 – *résultats de validation croisée*

297 séquences. Ces séquences sont identiques à celles du jeu d'apprentissage sauf dans certaines régions où elles comportent des insertions. Elles sont représentatives du type de données que PRHoD sera amené à traiter. Ce jeu de test a été enrichi de 297 contre-exemples également bruités puis classé par PRHoD (voir table 2). Le temps d'exécution sur un PC doté d'un processeur pentium 1,6 GHz à également été indiqué à titre indicatif. Les courbes ROC (*Receiver Operating Characteristic curve*) ont été utilisées pour évaluer la qualité du classement. Le score ROC représente l'intégrale normalisée de la courbe ROC, donc une mesure du taux de vrais positifs sur faux négatifs.

Classifieur	ROC	ROC50	temps d'exécution
Spectrum	**0,997**	0,984	6 s
Mismatch	**0,992**	0,957	4,5 min
HMotifs	**0,97**	0,97	6 s
Pairwise	**0,999**	0,999	14 min
LA kernel	**1**	1	23 min

TAB. 2 – *courbes ROC et temps d'éxécution sur des séquences bruitées*

5 Discussion et conclusion

Les performances des classifieurs en validation croisée (table 1) sont excellentes et permettent de conclure qu'ils sont aptes à découvrir des séquences inconnues. On observe qu'il y a deux types de classifieurs : ceux ayant une efficacité (pourcentage de séquence bien classées) proche de 100% (HMotifs, LA kernel et Pairwise) et ceux ayant des performances plus faibles (Spectrum et Mismatch). Cette différence peut s'expliquer par la flexibilité des classifieurs. Ceux travaillant sur les k-mots sont plus rigides dans leur reconnaissance de partie communes que HMotifs et que ceux basés sur le score SW.

Concernant les séquences bruitées, les intégrales de courbes ROC et ROC50 sont très proches de 1 (table 2), indiquant un classement presque parfait de ces séquences. Les classifieurs sont donc robustes au bruitage des données. La légère baisse d'efficacité de HMotifs s'explique par un artefact dans le classement des faux négatifs. Les classifieurs basés sur les k-mots, semblent donner de bons résultats mais cela est dû au fait que les séquences bruitées sont presque identiques aux séquences du jeu d'apprentissage. Ces performances ne sont donc pas indicatives

des résultats attendus sur des données nouvelles. En terme de temps d'exécution, les classifieurs basés sur les k-mots sont plus performant du fait de la simplicité de leurs algorithmes. Notons également que l'implémentation des classifieurs n'a pas été optimisée. Pour certains d'entre eux (LA kernel entre autre) il serait possible de réduire les temps de calculs.
La prochaine étape de ce travail sera de classer les cinq millions de séquences humaines avec les cinq classifieurs. Nous envisageons de définir des opérateurs d'agrégation pour obtenir un classement optimal des séquences.

Références

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, et D. J. Lipman (1997). Gapped blast and psi-blast : a new generation of protein database search programs. *Nucleic Acids Res 25*(17), 3389–3402.

Karplus, K., C. Barrett, et R. Hughey (1998). Hidden markov models for detecting remote protein homologies. *Bioinformatics 14*(10), 846–856.

Leslie, C., E. Eskin, et W. S. Noble (2002). The spectrum kernel : a string kernel for svm protein classification. *Pac Symp Biocomput*, 564–575.

Leslie, C. S., E. Eskin, A. Cohen, J. Weston, et W. S. Noble (2004). Mismatch string kernels for discriminative protein classification. *Bioinformatics 20*(4), 467–476.

Liao, L. et W. S. Noble (2003). Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships. *J Comput Biol 10*(6), 857–868.

Mikolajczak, J., G. Ramstein, et Y. Jacques (2004). Classification de protéines distantes par motifs hiérarchiques. In *Journées Ouvertes de Biologie, Informatique et Mathématique (JOBIM) 2005*.

Saigo, H., J. P. Vert, N. Ueda, et T. Akutsu (2004). Protein homology detection using string alignment kernels. *Bioinformatics 20*(11), 1682–1689.

Taylor, W. R. (1986). The classification of amino acid conservation. *J Theor Biol 119*(2), 205–218.

Vapnik, V. (1998). *The nature of statistical learning theory*. Springer-Verlag.

Summary

This paper presents PRHoD (Protein Remote Homology Detection), a tool aiming at finding new members of a protein family. PRHoD handles the whole discovery process, from data storage to identification of putative members. It comprises five classifiers, including an original one, based on Support Vector machines (SVM). They are especially designed for biological sequences. To discuss the performance of these classifiers in terms of accuracy and robustness, two test sets have been used, issued from protein and nucleotide data sources.

Alignement structural et classification hiérarchique pour l'extraction des cœurs structuraux

Khalid Benabdeslem*, Gilbert Deléage*
Christophe Geourjon*

*PBIL - IBCP-CNRS, 7 Passage du Vercors, 69367 Lyon Cedex 07, France
{kbenabde, c.geourjon, g.deleage}@ibcp.fr
http://pbil.ibcp.fr

Résumé. Dans ce papier, nous présentons une méthode originale de traitement de structures 3D des protéines. Cette méthode consiste à extraire des prototypes représentatifs de classes de protéines. Elle consiste dans un premier temps à calculer l'alignement structural en utilisant la méthode CE (Extension Combinatoire) sur les structures d'une même famille, et dans un deuxième temps, d'effectuer une classification ascendante hiérarchique (CAH) à partir de la matrice d'alignement. L'extraction des coeurs peut se faire à différents niveaux de l'arborescence de la CAH. Par conséquent, Chaque cœur extrait représente un prototype de la famille de structures qui lui est associée.

1 Introduction

L'extraction de connaissances à partir de données (ECD) est un domaine qui suscite depuis ces dernières années de nombreux travaux de recherche. En effet, ce domaine touche à de nombreux autres domaines de recherche, tels que les bases de données, les statistiques, l'intelligence artificielle ou la théorie de l'information, et est donc étudié sous différents aspects et dans de nombreux domaines d'applications. Parmi ces applications pratiques, celles liées à la bioinformatiques prennent une place de plus en plus importante.

Ce papier est consacré à ce type d'application en traitant particulièrement le problème difficile de la prédiction de la fonction tertiaire des protéines. De plus ce problème représente un véritable centre de préoccupation pour de nombreux chercheurs en bioinformatique. Plusieurs méthodes ont été développées dans cet intérêt (Marin, 2002 ; Przybylski 2004). Cependant, la plupart de ces méthodes trouvent des difficultés pour la détermination de la fonction de prédiction dans un espace de données à bas taux d'identité de séquence (moins de 25%). Pour cette raison, nous proposons une méthode d'extraction de connaissance à partir de données qui permet de mettre en avant les points communs entre les protéines structuralement similaires. Cette méthode consiste en l'extraction automatique des cœurs structuraux (que nous appelons ASCE). Etant basée sur l'alignement structural et la classification automatique, ASCE permet de fournir des cœurs structuraux à plusieurs niveaux hiérarchiques pour chaque classe de protéines.

2 L'alignement structural par la méthode CE (Extension combinatoire)

L'alignement structural s'est avéré un problème NP-Complet (Lathrop, 1994). Pour être simplifié ce problème nécessite l'application d'une variété d'heuristiques dont le choix est significativement important pour l'alignement. (Godzik, 1996) a montré que d'une part, plusieurs méthodes totalement différentes produisent différents alignements équivalents et d'autre part, la même méthodologie produisant le même score, fournit des alignements dans des positions différentes. Pour ces deux raisons, une bonne approche d'alignement structural est nécessaire.

L'algorithme de la méthode CE (Shindyalov, 1998) permet de construire un alignement entre deux structures de protéines. Cet algorithme nécessite une extension combinatoire d'un alignement défini par des pairs de fragments alignés (PFA) plutôt que des techniques conventionnelles comme la programmation dynamique et l'optimisation Monte Carlo.

2.1 Algorithme CE

2.1.1 Définition du chemin d'alignement

L'alignement entre deux structures de protéines A et B de longueurs respectives L_A et L_B, est considéré comme étant le plus long chemin P des PFA d'une taille m dans une matrice de similarité S de taille $(L_A\text{-}m) \times (L_B\text{-}m)$ qui représente toutes les PFA possibles et qui respectent le critère de similarité de structures. Un des critères suivants doit être satisfait pour chaque PFA i et i+1 dans le chemin d'alignement :

$$- P^{A}_{i+1} = P^{A}_{i} + m \text{ et } P^{B}_{i+1} = P^{B}_{i} + m \quad (1)$$

$$- P^{A}_{i+1} > P^{A}_{i} + m \text{ et } P^{B}_{i+1} = P^{B}_{i} + m \quad (2)$$

$$- P^{A}_{i+1} = P^{A}_{i} + m \text{ et } P^{B}_{i+1} > P^{B}_{i} + m \quad (3)$$

Où P^{A}_{i} est la position initiale du PFA de la protéine A dans la $i^{ème}$ position dans le chemin d'alignement (similairement pour P^{B}_{i})

La condition (1) décrit deux PFA alignées sans gaps et les conditions (2) et (3) représentent deux PFA avec gaps dans les protéines A et B respectivement.

2.1.2 Extension combinatoire du chemin d'alignement

Le chemin d'alignement est construit à partir d'un ensemble de PFA d'une taille fixe *m*. Deux fragments de taille *m,* respectivement de la première protéine et de la deuxième, forment une paire s'ils satisfont le critère de similarité décrit ci-dessous. La PFA initiale du chemin d'alignement peut être sélectionnée dans n'importe qu'elle position dans la matrice *S*. les PFA consécutives sont ainsi ajoutés suivant les condition (1), (2) ou (3). Pour limiter le nombre de gaps, deux conditions sont ajoutées (4) et (5) :

$$- P^{A}_{i+1} \leq P^{A}_{i} + m + G \quad (4)$$

$$- P^{B}_{i+1} \leq P^{B}_{i} + m + G \quad (5)$$

Où *G* représente et la taille maximale de gaps autorisés.

2.1.3 Des heuristiques d'évaluation de similarité et d'extension du chemin

Il existe plusieurs stratégies d'alignement qui diffèrent en complexité algorithmique. CE limite l'évaluation de la similarité aux mesures de distance suivantes :

(i) Distance D_{ij} utilisant un ensemble indépendant de distances inter - résidus

$$- D_{ij} = \frac{1}{m}\left(\begin{array}{l}\left|d^{A}_{p_i^A p_j^A} - d^{B}_{p_i^B p_j^B}\right| + \left|d_{p_i^A+m-1,\,p_j^A+m-1} - d_{p_i^B+m-1,\,p_j^B+m-1}\right| + \\ \sum_{k=1}^{m-2}\left|d^{A}_{p_i^A+k,\,p_j^A+m-1-k} - d^{B}_{p_i^B+k,\,p_j^B+m-1-k}\right|\end{array}\right) \quad (6)$$

(ii) Distance D_{ij} calculée en utilisant un ensemble complet de distances inter – résidus évaluées à :

$$- D_{ij} = \frac{1}{m^2}\left(\sum_{k=0}^{m-1}\sum_{l=0}^{m-1}\left|d^{A}_{p_i^A+k,\,p_j^A+l} - d^{B}_{p_i^B+k,\,p_j^B+l}\right|\right) \quad (7)$$

(iii) L'erreur quadratique (Rmsd) obtenue des structures superposées de manière optimale (Hendrickson, 1979) où :

- D_{ij} représente cette fois-ci la distance entre deux combinaisons de deux fragments des deux protéines *A* et *B* définies par deux PFA dans les positions *i* et *j* dans le chemin d'alignement où $i \neq j$. Dans le cas d'une PFA singulière, $i = j$ dans D_{ij}.
- d^{A}_{ij} représente la distance entre les résidus *i* et *j* dans la protéine *A* (similairement pour d^{B}_{ij})
- *m* représente la taille des fragments.

La mesure de distance (i) est utilisée pour évaluer la combinaison des deux PFA, la mesure (ii) est utilisée pour la PFA singulière et la mesure (iii) est utilisée en dernière étape pour sélectionner quelques autres alignements en optimisant le nombre de gaps.

3 Classification des protéines pour l'extraction des cœurs structuraux

Dans cette section, nous montrons comment extraire des cœurs structuraux à partir des classes de protéines d'une même famille. Pour commencer nous nous basons sur une classification de protéines en terme de topologie proposée par CATH (Orengo et al, 2004).

Ensuite pour chaque famille de protéine, nous appliquons la méthode CE décrite dans la section 2. Nous obtiendrons donc une matrice de dissimilarités entre les structures de la famille sur la métrique Rmsd (la mesure iii dans la section précédente). Cette matrice est ainsi présentée à l'algorithme de la CAH. Nous obtiendrons enfin, un dendrogramme par famille.

Nous proposons de montrer un exemple d'application des étapes de la méthode ASCE sur la famille CATH : 1.50.10
(cf :http://cathwww.biochem.ucl.ac.uk/latest/class1/50/10/index.html). Cette famille contient cinq structures représentées par les codes suivants : 1dl2A0, 1gxmA0, 1kwfA0, 1n7oA0 et 1qazA0. Chaque structure possède un fichier « PDB » (Protein Data Bank) qui représente les coordonnées 3D (x, y, z) de chaque acide aminée (symbole) constituant la séquence protéique (FIG1).

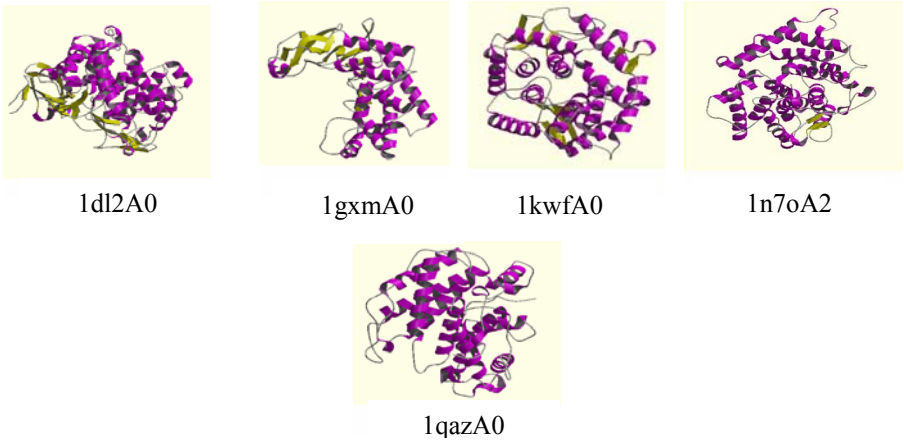

FIG. 1 – *Les structures 3D de la famille CATH: 1.50.10*

L'application de CE entre chaque paire de structure3D (pratiquement entre leurs fichiers PDB) de FIG.2 fournit une matrice de dissimilarités (en terme de Rmsd) de taille 5×5. Cette matrice est ainsi présentée à l'algorithme de la CAH qui génère une taxonomie regroupant les structures à plusieurs niveaux hiérarchiques (FIG.2).

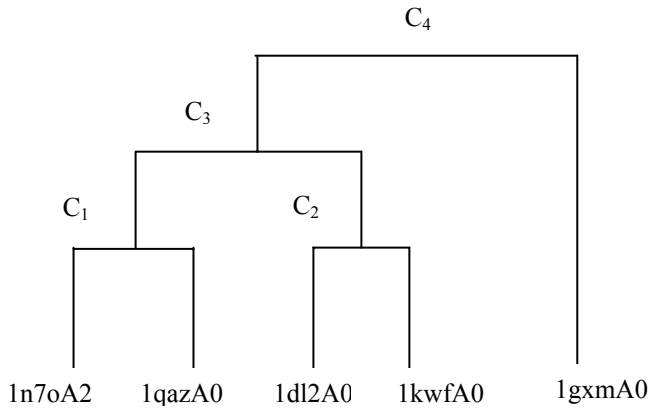

FIG. 2 –*Un dendogramme représentant la famille 1.50.10*

Dans FIG.2, les feuillets représentent les structures et chaque nœud dans le dendogramme définit un cœur structural. L'utilisateur pourra donc choisir d'extraire des cœurs structuraux à tout niveau dans la hiérarchie.

Exemple :

Pour extraire le cœur structural C_3, on doit sélectionner deux structures à partir des ensembles {1n7oA, 1qazA0} et {1dl2A0, 1kwfA0}. Nous montrons dans la section suivante comment élire ces deux structures.

3.1 Election des structures composants le cœur structural

Le cœur structural est un ensemble de blocs d'acides aminés (symboles constituant les séquences) qui couvrent au « mieux » l'ensemble des structures de la famille. Il représente le résultat de l'alignement structural des deux structures les plus représentatives de la famille.

Pour élire ces deux structures nous procédons comme suit :

- Si le nœud couvre deux protéines, nous considérons ces deux structures comme étant élues (cf. C_1 et C_2 dans FIG.2).
- Si le nœud est associé à une classe contenant plus de deux structures (cf. C_3 ou C_4 dans FIG.2), les deux structures élues seront extraites des deux sous classes formant la classe du nœud. Ce sont les deux structures les plus similaires en 3D des deux sous classes (FIG.3)

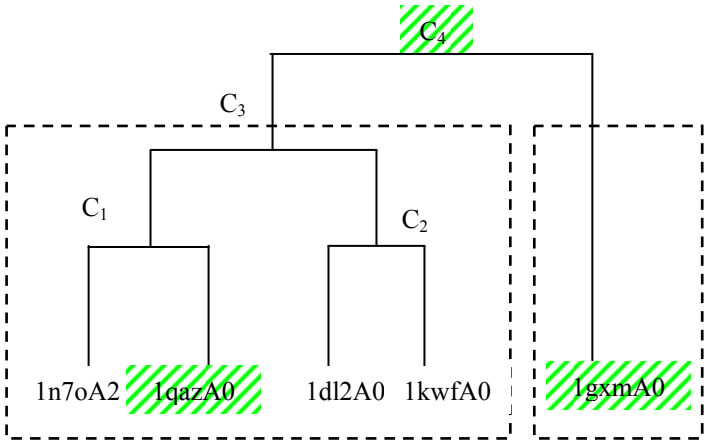

FIG. 3 – *Election des structures pour les cœurs*

Dans FIG3. Pour élire les structures qui construisent le cœur C_4, nous considérons les deux sous classes du nœud associé au cœur (sous classes C_3 et {1gxmA0}). Nous sélectionnons par la suite dans ces deux sous classes, les deux structures les plus structuralement similaires (1qazA0 et 1gxmA0 selon CE).

3.2 Extraction des coeurs

Les structures ainsi élues sont présentées à nouveau à l'alignement structural CE. Un alignement de séquences en est donc déduit. Il représente une série de bloc d'acides aminés des deux élues.

Nous appliquons un filtrage basé sur le calcul de Rmsd local entre chaque paire d'acides aminés des deux alignements des deux structures élues. Le seuil étant fixé au Rmsd global issu du CE. En d'autres termes, pour être préservés, deux acides aminés doivent afficher un Rmsd qui ne dépasse passe pas le Rmsd global de l'alignement des deux protéines

Exemple :

1gxmA0

MTGRMLTLDGNPAANWLNNARTKWSASRADVVLSYQQNNGGWPKNLDYNSVGNGGGGNESGTIDNGATITEMVFLAE
VYKSGGNTKYRDAVRKAANFLVNSQYSTGALPQFYPLKGGYSDHATFNDNGMAYALTVLDFAANKRAPFDTDVFSDND
RTRFKTAVTKGTDYILKAQWKQNGVLTVWCAQHGALDYQPKKARAYELESLSGSESVGVLAFLMTQPQTAEIEQAVRAG
VAWFNSPRTYLEGYTYDSSLAATNPIVPRAGSKMWYRFYDLNTNRGFFSDRDGSKFYDITQMSLERRTGYSWGGNYGTSII
NFAQKVGYL

1qazA0

GSHPFDQAVVKDPTASYVDVKARRTFLQSGQLDDRLKAALPKEYDCTTEATPNPQQGEMVIPRRYLSGNHGPVNPDYEPV
VTLYRDFEKISATLGNLYVATGKPVYATCLLNMLDKWAKADALLNYDPKSQSWYQVEWSAATAAFALSTMMAEPNVDT
AQRERVVKWLNRVARHQTSFPGGDTSCCNNHSYWRGQEATIIGVISKDDELFRWGLGRYVQAMGLINEDGSFVHEMTRH
EQSLHYQNYAMLPLTMIAETASRQGIDLYAYKENGRDIHSARKFVFAAVKNPDLIKKYASEPQDTRAFKPGRGDLNWIEYQ
RARFGFADELGFMTVPIFDPRTGGSATLLAYKP

Composante1 du cœur

---.....RM--------A---------.....W---RAD-------N-GG----------.....T-D--.ATITEMV.F--E--..K-------R--.V..R-AAN--V--
.Q--T--------------HA.T.F---G.MAYALT.V--F---KR---D-DV-SDNDRTRFKTAVTKGTDYILKA-W----------L-..G-
-S--VLA-LM-.Q...P.QT--I.E-AV.R-GVAWFNS--TYLEG----------KMW..Y---------G-F----------I-------R---SWGG-.Y---I--.F-
-.....K----.....

Composante 2 du coeur

--------VV--------Y----------P---KEY-------A-TP----------Y-R---KISATLG-L--Y----K-------V---C--N-
MLD--A---A--L--------------SW-Q-E---W-AATAAF-L--S---MA---E-NV-DTAQRERVVKWLNRVARHQTSF-P----------
--------S---N--Y--RGQ-AT--G---K-DE--L-R-WG-G-RYVQAMG--LINED----------GSF--E---------T-R----------H-------E---
HYQN--M---L---A--------D----------

L'exemple ci – dessus montre les deux structures (1gxmA0 et 1qzA0) qui sont élues pour la construction du cœur C_4 après avoir appliqué le filtre sur les Rmsd inter – résidus. Ce cœur couvre 100% de la population de la famille 1.50.10 avec 34,5% de préservation de résidus dans les composantes du cœur (FIG. 4).

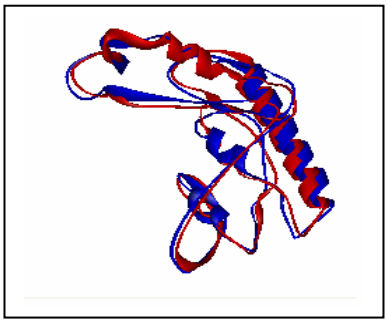

FIG. 4 – *Représentation 3D du cœur C_4*

FIG. 4 montre les deux composantes constituant le cœur structural C_4 pour toute la famille 1.50.10 (C_4 étant situé dans la racine de la taxonomie). La partie bleue représente les acides aminés conservés dans 1gxmA0 et la partie rouge représente les acides aminés conservés dans 1qzA0.

Dans les composantes du cœurs ci – dessus, les '-' représentent les gaps issues de l'alignement structural entre 1gxmA0 et 1qzA0 et les '.' Représentent les acides aminés qui ne sont pas conservés dans les deux structures (et qui ne font donc pas partie du cœur structural).

4 Alignement des protéines sur les cœurs structuraux pour la reconnaissance séquence – famille de structures.

Nous avons choisi une base de structures dans CATH, qui contient 86 familles. Ensuite, nous avons calculé l'identité de séquence en utilisant FASTA (Pearson, 1990) entre chaque paire de séquences de la base. Le pourcentage moyen d'identité était de 20%.

En contre partie, nous avons aligné chaque séquence de la base par rapport au cœur de sa famille par la méthode ClustalW (Thompson et al, 1994). Nous avons appliqué le même procédé de calcul d'identité et le résultat est passé à 30% d'identité.

Ce résultat nous montre bien que l'alignement de la base sur les cœurs a fait augmenté l'identité de séquence. Nous sommes donc, passé d'un espace de données non significatif (l'espace de données brutes : espace de séquences) à un espace de données à taux d'identité amélioré (espace de Meta données : espace de séquences alignées).

Par ailleurs, nous avons calculé la similarité entre les données brutes et les cœurs structuraux , nous l'avons comparé par rapport au classement proposé par CATH, le résultat était de l'ordre de 30% de reconnaissance séquence – famille.

Nous avons ensuite appliqué le même processus en remplaçant la base des séquences brutes par la base des séquences alignées sur les cœurs, le pourcentage était donc de 34% (nous rappelons que ces pourcentages sont faibles à cause de l'hypothèse du départ sur l'identité de séquence qui est assez faible dans l'ensemble de la base et qu'aucun modèle de reconnaissance n'a été construit, nous nous sommes basés que sur le calcul naïf de similarité entre les séquences et les coeurs). Cependant, nous constatons que l'extraction des cœurs structuraux représente un facteur déterminant pour les systèmes de reconnaissance de repliements.

5 Conclusion

Le travail présenté dans cet article, représente une étape d'amélioration de tout système de reconnaissance de repliements voire la prédiction tertiaire des protéines. Nous avons exploité des informations pertinentes qui pouvaient être extraites à partir d'une classification de protéines. Grâce à l'alignement structural proposé par CE et une classification faite par CAH, nous avons pu établir pour chaque famille de protéines un représentant significatif qui couvre l'ensemble des résidus des protéines de la famille et qui sert à aligner au mieux toutes ces protéines. En perspective, ce travail permet de normaliser la base de séquences qui pourrait être utilisée comme base d'apprentissage dans un système de reconnaissance ou de modélisation 3D.

Références

Bouroche J-M et Saporta G. (1994) *L'analyse des données*. Presse universitaire de France.

Godzik, A. (1996). *The structural alignment between two proteins: Is there a unique answer?* Protein Science, 5, 1325-1338.

Hendrickson, W.A. (1979) *Transformations to optimize the superposition of similarstructures.* Acta Cryst., A35, 158-163.

Lathrop, R.H. (1994). *The protein threading problem with sequence amino acid interaction preferences is NP-complete*. Prot. Engng, 7, 1059-1068.

Lechevallier Y et al (1996). *Statistiques et Méthodes Neuronales*, Dunod.

Marin A, Pothier J, Zimmermann K, Gibrat JF. (2002). *FROST: a filter-based fold recognition method*. Proteins ;49(4):493-509.

Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., et Thornton, J.M. (1997) CATH- *A Hierarchic Classification of Protein Domain Structures*. Structure. **5**, 1093-1108. (http://www.biochem.ucl.ac.uk/bsm/cath/)

Pearson W. R. (1990) *Rapid and Sensitive Sequence Comparison with FASTP and FASTA, Methods* in Enzymology **183**, 63 - 98

Przybylski D et Rost B. (2004). *Improving fold recognition without folds*. J Mol Biol.;341(1):pp255-69.

Shindyalov I.N et Bourne P.E. (1998) *Protein structure alignment by incremental combinatorial extension (CE) of the optimal path*. Protein Engineering **11**, 739-747. (http://cl.sdsc.edu/ce.html)

Thompson JD, Higgins DG & Gibson TJ (1994) *CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res*, **22**, 4673-4680

Remerciements

Ce travail a été réalisé dans le cadre du projet Genoto3D soutenu par l'ACI « Masse de données » que nous tenons à remercier pour le soutien financier et scientifique en regroupant plusieurs laboratoires réunissant des chercheurs en informatique, bioinformatique et biologie.

Summary

In this paper, we present an original approach of treating 3-dimentional structures of proteins. This approach consists on extracting representative prototypes from classes of proteins. In the first step, we calculate a structural alignment for each class (family) of structures using the combinational extension (CE) alignment methodology. In the second step, we carry out from each alignment's matrix, a taxonomy created with ascendant hierarchical clustering (AHC). Finally, structural cores are extracted from all families and each core represents a prototype for each class of structures.

Construire des dictionnaires pour l'identification d'entités biologiques complexes

Julien Lorec[*,**], Gérard Ramstein[**], Yannick Jacques[*]

[*]INSERM U601, Département de Cancérologie, équipe 3: cytokines et récepteurs
{julien.lorec,yjacques}@nantes.inserm.fr
[**]LINA, équipe C.O.D, école polytechnique de l'université de Nantes
gerard.ramstein@polytech.univ-nantes.fr

Résumé. Nous proposons, dans cet article, une méthodologie de construction de dictionnaires contrôlés pour l'extraction d'information textuelle en biologie moléculaire. Nous nous limitons ici volontairement à définir des dictionnaires servant à identifier le nom des gènes ou des protéines, des souches cellulaires, des expériences ou protocoles expérimentaux, des sites de liaison aux facteurs de transcription et des facteurs de transcription dans les publications scientifiques traitant de la biologie humaine. Nous allons décrire d'une part les diverses sources nécessaires à leur création et d'autre part nous nous efforcerons d'expliquer les différentes techniques mises en place afin de répondre aux difficultés spécifiques de la nomenclature du domaine. La couverture relative de nos dictionnaires, après expertise du contenu, est évaluée à partir du corpus de référence GENIA.

1 Introduction

La détection et la reconnaissance de termes correspondant à des objets et des concepts biologiques est la première étape, cruciale, qui précède l'utilisation de l'information contenue dans les textes biomédicaux.
Néanmoins, cette étape se révèle particulièrement ardue dans les publications scientifiques de spécialité de biologie moléculaire. Certains auteurs ne respectent pas ou peu les conventions, ces dernières n'étant d'ailleurs que trop souvent consultatives et non obligatoires. Par ailleurs, la nomenclature du domaine n'est pas standardisée et est peu structurée. Un bon aperçu des différents problèmes liés à la nomenclature en biologie moléculaire est synthétisé dans (Tuason et al. (2004)). Les systèmes d'extraction d'information en biologie moléculaire doivent être capables de gérer à la fois :

- la présence de termes synonymes et la résolution des différentes abréviations et acronymes,
- la variabilité des mots tant au niveau de l'orthographe que de la morphologie et de la syntaxe mais aussi d'un point de vue lexico-sémantique et de la présence d'insertions/déletions et permutations,
- la présence de noms ambigus que se soit entre des entités de même nature, entre des entités de natures différentes ou des collisions avec le dictionnaire anglais standard.

On peut ditinguer les méthodes d'identification des entités biologiques selon qu'elles utilisent des dictionnaires ou non. Néanmoins, ces deux types d'approches ne répondent pas aux mêmes questions. Les méthodes qui n'emploient pas de dictionnaire (Collier et al. (2000), Fukuda et al. (1998)) sont extrêmement utiles pour en construire. Elles peuvent en effet découvrir de nouvelles entités biologiques potentielles qui seraient absentes des dictionnaires. En revanche, ces techniques sont très peu adaptées aux tâches relatives à l'exploration et à la description des relations entre entités biologiques. Notamment, les problèmes de synonymie et d'homonymie des noms des entités de biologie moléculaire ne peuvent être gérés à ce niveau et nécessitent des connaissances *a priori*. En comparaison, les méthodes qui dépendent de l'utilisation de dictionnaires (Fundel et al. (2005), Koike et Takagi (2004)) sont à même de répondre à ces mêmes difficultés et peuvent être intégrées dans des systèmes d'extraction d'information de haut niveau.
Nous présentons ici une méthodologie de création de dictionnaires qui se veut assez générique pour gérer simultanément divers descripteurs biologiques de nature variée chez l'homme. Les différentes techniques exposées tenteront de répondre aux mieux aux problèmes liés au manque de standardisation dans le domaine. Leur couverture sera étudiée en utilisant le corpus de référence GENIA (Jin-Dong et al. (2003)).

2 Ressources utilisées

Nous nous intéressons plus particulièrement à la découverte des objets biologiques suivants dans les textes : d'une part les *gènes et protéines humaines*, certaines zones particulières de l'ADN (les *sites de liaison aux facteurs de transcription*) humain, les *facteurs de transcription* humains, les diffèrentes *souches de cellules* humaines, les *tissus* et *organes* humains mais encore les *protocoles expérimentaux et techniques, les appareillages utilisés au cours d'expériences biologiques*. Nous avons sélectionné un nombre restreint mais complet et correctement expertisé de bases de données ou d'ontologies publiques afin de construire nos différents dictionnaires. LocusLink[1], HUGO[2], GDB[3] et OMIM[4] ont servi à référencer les gènes et protéines, TRRDSITE[5] les sites de liaison aux facteurs de transcrition, TFD[6], COMPEL[7], TRRDFACTORS et TFFACTOR[8] les facteurs de transcription et certaines sources du MetaThesaurus UMLS[9](Lindberg et al. (1993)) les cellules, tissus, organes, protocoles expérimentaux et techniques et appareillages.

Ces bases de données proposent pour chaque entrée à la fois un ou plusieurs noms complets usuels ainsi qu'un ou plusieurs symboles/acronymes/abréviations officiels ou non définissant un même objet biologique. Ces différents alias peuvent être encore en usage ou ne plus avoir

[1]http ://www.ncbi.nlm.nih.gov/LocusLink/
[2]http ://www.gene.ucl.ac.uk/nomenclature/
[3]http ://www.gdb.org/
[4]http ://www.ncbi.nlm.nih.gov/entrez/query.fcgi ?db=OMIM
[5]http ://wwwmgs.bionet.nsc.ru/mgs/gnw/trrd/
[6]gopher ://gopher.nih.gov/77/gopherlib/indices/tfd/index
[7]http ://compel.bionet.nsc.ru/new/index.html
[8]http ://www.gene-regulation.com/pub/databases.html#transfac
[9]CRISP Thesaurus 2003, Gene Ontology 2002_12_16, Medical Subject Headings 2004_2003_08_08, NCBI Taxonomy 2003, NCI Thesaurus 2001a, Standard Product Nomenclature 2002 et UMDNS : product category thesaurus 2003

avoir cours dans la littérature contemporaine. Les différentes entrées sont fusionnées grâce aux symboles HUGO communs pour les gènes et protéines. Malheureusement, pour les facteurs de transcription, les identifiants répertoriés sont uniques à chaque base de données et ne peuvent être directement utilisées pour les relier entre elles. Néanmoins, les facteurs de transcription appartiennent à un sous-ensemble de la famille biologique des gènes et des protéines. Aussi nous fusionnons les entrées relatives aux gènes et protéines et aux facteurs de transcription lorsqu'au moins un alias est commun. Les alias ainsi mis en commun ne sont plus ceux d'un gène ou d'une protéine mais ceux d'un facteur de transcription. Les entrées conflictuelles ou orphelines sont examinées manuellement. Nous ne nous intéressons qu'aux entités d'origine humaine ou à défaut de mammifère. Nous utilisons pour cela les informations contenues dans les champs relatifs à l'espèce ou à la classe dans les bases de données multi-espèces.
Contrairement à d'autres approches (Koike et Takagi (2004)) les différents objets manipulés dans les dictionnaires n'entretiennent entre eux aucune relation d'appartenance ni de composition. L'exception notable étant la classe des facteurs de transcription. Néanmoins l'intégration d'informations hiérarchiques au sein de nos dictionnaires est envisageable grâce à l'utilisation d'une part des classifications issues de SWISS-PROT, PIR et INTERPRO pour les gènes et protéines et facteurs de transcription et d'autre part ceci[10] pour les sites de liaison aux facteurs de transcription et des ressources UMLS.

3 Composition et description

3.1 Génération de variants

Nous stockons dans nos dictionnaires les différentes représentations lexico-sémantiques, grammaticales et syntaxiques de chaque objet biologique des différentes bases de données. L'intérêt ici est double : la génération d'alias alternatifs est réalisée d'une part à partir d'informations inaccessibles lors de l'étape de détection des entités nommées dans les textes et d'autre part elle peut être relativement coûteuse en temps de calcul.
Pour l'ensemble des bases de données à l'exception du Metathesaurus, chaque tuple peut contenir plusieurs alias d'une même entité. Le format de ces entrées suivent les recommandations du comité HUGO (avec quelques nuances) :

1. les synonymes décrivant la même entité sont séparés par des points-virgules (ex "stem cell growth factor ; lymphocyte secreted C-type lectin"),
2. les synonymes peuvent être complétés par de termes descriptifs, séparés par des virgules (ex "sodium channel, voltage-gated, type XI, alpha"),
3. les termes entre parenthèses peuvent à la fois être le nom d'autres espèces où a été découverte l'entité ou un nom alternatif complet (ex "SCO (cytochrome oxidase deficient, yeast) homolog 1"). Les séparateurs points-virgules et virgules gardent leurs rôles respectifs au sein des parenthèses.

Chaque alias est alors considéré comme une forme variante valide d'une même entité et constitue une entrée à part entière.
Nous utilisons aussi les informations issues des nomenclatures spécifiques de chaque base de données afin de générer des formes inédites qui peuvent être retrouvées uniquement dans les

[10] http ://www.gene-regulation.com/pub/databases/transfac/cl.html

publications scientifiques.
Des alias sont générés *de novo* en utilisant d'une part les combinaisons issues des arrangements des termes descriptifs (ex "aconitase 1 soluble", "soluble aconitase 1" et "aconitase 1") et d'autre part les différentes formes mixtes acronymes/noms complets (ex "chemokin like receptor 1", "CMKLR1", "CMKL receptor 1", "CMK light receptor 1", "chemokin like R 1" et "chemokine LR 1") d'une même entité. Lorsque nous disposons à la fois d'un acronyme et d'une forme complète du nom d'une entité, nous essayons de développer les lettres et combinaisons de lettres de l'acronyme à partir d'un lexique contrôlé si et seulement si le terme développé est présent dans le nom complet et à une position compatible à la fois dans l'acronyme et dans le nom complet.

3.2 Normalisation

Autant nous stockions chaque synonyme d'une même entité dans nos dictionnaires, autant il est inutile de s'embarasser des différentes formes orthographiques et morphologiques de chacun.
Beaucoup de noms d'entités en biologie peuvent inclure des structures sujets - verbes - compléments. Afin de réduire la complexité de tels noms, chaque forme n'est conservée que sous la forme de noms composés (*compound nouns*) et si elle n'existe pas dans les bases de données, elle sera alors générée. Ainsi, par exemple, "linker for activation of T cells", "activation of T cells linker" et "T cells activation linker" n'est gardée dans le dictionnaire que sous sa dernière forme. Le problème des majuscules et minuscules en biologie est très important surtout pour le nom des molécules et des cellules (par exemple "cAMP" ne correspond pas à la même entité que "CAMP"). La présence de caractère spéciaux et d'espace est souvent aussi source d'ambiguïté (par exemple "Il2R", "IL 2R" et "Il2-R" représentent la même entité). Chaque mot est alors découpé à chacun de ses points de rupture de casse. Les suites consécutives de lettres en minuscules, de lettres en majuscules, de caractères spéciaux et numéros sont séparées par un espace. Une exception subsiste : un caractère majuscule qui débute la chaine ou qui est précédé par un caractère non-alphanumérique et est suivi par une minuscule n'est pas isolé. Dans un deuxième temps, toute la chaîne de caractères est passée en minuscule et les lettres non-alphanumériques sont remplacées par un espace. Par exemple, "cAMP", "c-Amp" et "c Amp" se réfèrent à la même entité "c amp" dans notre dictionnaire et est différent de "CAMP" présent lui sous la forme "camp". Quelques contre-exemples doivent être traîtés à part tels que les abréviations usuelles de l'unité de mesure kilo-Dalton qui apparaissent soit sous la forme "kD", "KDa" ou "kd" et les caractères "-" et "." utilisés en combinaison avec des numéros. En principe, les chiffres romains sont aussi remplacés par leurs équivalents numériques arabes et les symboles grecs sont écrits en toute lettre. Quelques exceptions demeurent encore, tel que le caractère "X" qui, précédé d'une instance du mot "chromosome" ou suivi du mot "ray", ne désigne pas dans ces cas de figure particuliers le numéro 10. Finalement, chaque variant qui ne correspond pas à un acronyme est lemmatisé en utilisant l'algorithme de Porter afin de s'afranchir, en particulier, des formes plurielles et de la différence entre les suffixes d'origine américaine ou britannique. Une lemmatisation forte à comme principal désavantage d'occulter les formes actives et passives des verbes (par exemple "neutrophil activated peptide" et "neutrophil activating peptide" ne peuvent représenter la même entité biologique).

4 Résultats et discussion

Les bases de données utilisées contiennent de trop nombreuses erreurs de nomenclature et de noms inappropriés. Afin de conserver une fiabilité maximale, nous supprimons de manière automatique les termes appartenant à des concepts de plus haut niveau (ex "transmembrane transporter"), les termes sans intérêt (ex "uncharacterized protein") ou parasites (ex "contains only BH3 domain", "antigen identified by monoclonal antibody", "entry checked") à partir de lexiques contrôlés et de règles. Une expertise manuelle poussée est ensuite réalisée.
La mesure de la couverture des dictionnaires sur le corpus GENIA présentée sur le tableau 1 à été réalisée grâce à la méthode exposée dans (Lorec et al. (2006)). Seul un sous-ensemble du corpus à été analysé afin, d'une part, de s'assurer de la qualité de l'appariement et du recoupement de nos classes d'objets biologiques avec celles présentes dans GENIA, et d'autre part, dans le but d'ajouter notre propre annotation lorsque le concept n'était pas préalablement référencé. Dans nos résultats, nous avons comptabilisé en tant que vrais positifs les objets biologiques correctement détectés mais originaires de mammifères et non plus spécifiquement humains. Ce problème ne peut être adressé que si l'on considère le contexte de l'utilisation de l'entité dans le texte et nous avons préféré pour le moment ignorer cette difficulté.

TAB. 1 – *Composition des dictionnaires et couverture sur GENIA*

	G	S	F	C	O	P	TOTAL
Nombre de variants	183196	6524	11379	508	1768	1284	204659
Nombre d'entités	43259	2268	1773	312	1202	769	49583
Précision	0.88	0.2	0.85	0.92	0.8	0.9	0.83
Rappel	0.81	0.33	0.89	0.77	0.88	0.75	0.81
Nombre de vrais positifs	44	2	34	37	32	18	167
Nombre de faux positifs	6	8	6	3	8	2	33
Nombre de faux négatifs	10	4	4	11	4	6	39

G = gènes et protéines, S = sites de liaison aux facteurs de transcription, F = facteurs de transcription, C = cellules, O = tissus ou organes, P = protocoles et appareillages d'expérience

Les principales sources de faux positifs sont liées à des problèmes de désambiguation. D'une part certaines entités identifiées correspondent au nom de la famille et non au membre exact (par exemple, "interleukin 1 receptor accessory protein" est absent des dictionnaires et est associé par défaut à l'entité "interleukin 1 receptor"). D'autre part, et si l'on ne tient pas compte du mésappariement des espèces aux entités identifiées, la seconde difficulté consiste dans l'effort de nettoyage des dictionnaires qu'il reste encore à produire. La principale source de faux négatifs est l'absence de la forme variante correspondante dans nos dictionnaires ou parce que l'entité n'a pas du tout été référencée à l'origine dans les bases de données. Les raisons majoritaires des mauvais résultats obtenus avec les sites de liaison aux facteurs de transcription sont des noms très courts et peu remarquables (ex "A site"), très souvent identiques à ceux des facteurs de transcription.

5 Conclusion

Cet article présente un procédé simple et relativement générique de création de dictionnaires de noms d'objets biologiques nécessaires à leur identification au sein de publications

scientifiques du domaine de la biologie moléculaire. Le principal facteur influant sur la qualité de tels dictionnaires est le nettoyage systématique et expertisé des données redondantes, inappropriées ou erronées apportées par l'automatisation de la tâche. Néanmoins plusieurs questions restent en suspens. Les noms des entités répertoriées dans les bases de données sont très souvent formels et très descriptifs. On a rarement dans les textes la totalité des termes définis dans les bases de données (ex "precursor", "type" ou "member"). Ce problème est en cours d'examination. Le fait de stocker les formes variantes des entités en minuscule rique d'introduire des collisions avec le dictionnaire anglais standard au niveau de l'étape de détection des entités nommées. (Fundel et al. (2005)) utilisent un classifieur SVM pour tenter de pallier à ce problème chez la drospohile, avec de très bons résultats.

Références

Collier, N., C. No, et J. Tsujii (2000). Extracting the names of genes and gene products with a hidden markov model. *Proc. COLING 2000*, 201–207.

Fukuda, K., T. Tsunoda, A. Tamura, et T. Takagi (1998). Toward information extraction : Identifying protein names from biological papers. *Proc. of the Pacific Symposium on Biocomputing '98*.

Fundel, K., D. Güntler, R. Zimmer, et J. Apostolakis (2005). A simple approach for protein name identification : prospects and limits. *BMC Bioinformatics*, 6(Suppl. 1) :S15.

Jin-Dong, K., T. Ohta, Y. Teteisi, et J. Tsujii (2003). Genia corpus - a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl. 1) ;i180–i182.

Koike, A. et T. Takagi (2004). Gene/protein/family name recognition in biomedical literature. *Proceedings of HLT/NAACL BioLINK workshop*, 9–16.

Lindberg, D., B. Humphreys, et A. McCray (1993). The unified medical language system. *Methods Inf Med.*, 32(4) :281–91.

Lorec, J., G. Ramstein, et Y. Jacques (2006). Extraction et identification d'entités complexes à partir de textes biomédicaux. *Extraction et gestion des connaissances*, RNTI–E–3, à paraître.

Tuason, O., L. Chen, H. Liu, J. Blake, et C. Friedman (2004). Biological nomenclatures : Source of lexical knowledge and ambiguity. *Proceedings of the Pacific Symposium of Biocomputing*, 9 :238–249.

Summary

We propose a methodology to construct dictionaries for IE in the fields of molecular biology. In this paper, we intentionally limit ourselves to the following entities: human genes or proteins, cellular lines, experiments or biological protocols, transcription factor binding sites and transcription factors. We will first describe the differents sources gathered in order to create such dictionaries and then the techniques we have developed to overcome the specific difficulties encountered in the molecular biology nomenclature. The GENIA corpus has been used to evaluate the relative coverage of these dictionaries, with extensive curation.

Extraction et Sélection des n-grammes pour le Classement des Protéines

Faouzi Mhamdi*, Mourad Elloumi*, Ricco Rakotomalala**

*URPAH, FST, Université d'El Manar, Campus Universitère 1060 Tunis, Tunisie
Faouzi.mhamdi@ensi.rnu.tn, Mourad.Elloumi@fsegt.rnu.tn
**Labo ERIC, Université Lyon2, Lyon, France
Ricco.rakotomalala@univ-lyon2.fr

Résumé. Le classement de protéines est une activité importante pour le biologiste. Avec l'augmentation croissante de la taille des banques de données, le classement manuel est devenu impossible, il est donc nécessaire de mettre en œuvre des stratégies informatiques pour automatiser le processus. Nous présentons un cadre global inspiré de la démarche de l'extraction de connaissances à partir de données pour classer automatiquement les protéines à partir de leurs structures primaires. Ce cadre comporte deux grandes étapes : la première, la phase de pré-traitement, consiste à extraire des descripteurs à partir de la description originelle des données, nous avons utilisé la technique des n-grammes bien connue dans la catégorisation de textes ; la seconde, la phase d'apprentissage, consiste à utiliser les techniques d'apprentissage, intégrant la phase de sélection de variables, pour classer automatiquement les protéines. Les expérimentations sur banques de données réelles permettent d'obtenir des résultats encourageants. Cet article fait le point sur les différentes alternatives que nous avons analysées à chaque étape du processus de classement.

1 Introduction

Ces dernières années, les données biologiques n'ont cessé de se multiplier d'une façon spectaculaire. Elles sont rassemblées dans des banques de données publiées sur Internet, elles sont principalement représentées par des séquences d'ADN, d'ARN, de protéines, etc Gibas et Jambeck (2002). Ces banques sont créées et complétées par des biologistes après des travaux de séquençage et d'expérimentation dans les "payasses" (in vitro). Pour cela il y a un besoin en informatique pour modéliser le stockage, la conservation des données, la possibilité de lancer des requêtes efficacement ; au-delà de cet aspect purement organisationnel, on se rend compte qu'il y a sûrement des régularités et des informations intéressantes dans ces données, dans ce cas, on est plutôt dans le domaine de la fouille de données. Ce travail a pu être formalisé dans un cadre générique : l'extraction de connaissance à partir des données (ECD, KDD en anglais - Knowledge Descovery in Database) Fayyad et al. (1996). L'ECD est un processus qui se compose de trois grandes étapes : a) prétraitement de données, b) apprentissage ou fouille de données, c) post-traitement de données. La figure 1 présente le processus. Le prétraitement de données consiste à rassembler, sélectionner, nettoyer et filtrer ces données. La deuxième

phase est considérée comme le cIJur du processus.. Il s'agit d'appliquer les techniques de fouille de données sur les données préparées dans le but d'en extraire de la connaissance. Ces connaissances peuvent être présentées par une classification, prédiction, etc. Enfin, la phase de post-traitement consiste à valider, visualiser et utiliser ces connaissances. Le processus d'ECD

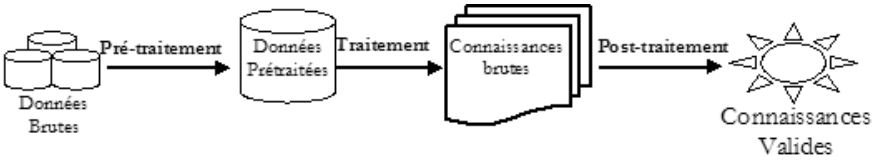

FIG. 1 – *Processus ECD*

a montré son efficacité dans les domaines du marketing, des géosciences, etc. Ces domaines ont pour point commun le grand volume de données à gérer. Les scientifiques ont essayé d'adapter ce processus pour qu'il supporte les données biologiques. Maintenant on parle de Bio-data mining (BioDM). Le BioDM est la science d'extraction de connaissances à partir des données biologiques en appliquant des techniques de fouille de données. Plusieurs travaux de recherche sont réalisés dans ce cadre. Les principales applications sont la classification des séquences biologiques, la prédiction des structures protéiques, la prédiction des gènes, etc.

Une protéine possède différentes structures de représentation. On trouve la structure primaire, la structure secondaire et la structure tertiaire (3D). Plusieurs problématiques en liaison avec les protéines ont été identifiées, telles que la prédiction des structures protéiques, alignement des séquences protéiques, prédiction des fonctions de protéines et essentiellement le classement des protéines.

Notre travail consiste à développer une nouvelle approche de classement des protéines basée sur leurs structures primaires. Cette approche est inspirée du processus d'ECD. Une protéine peut être décrite par sa structure primaire, il s'agit d'une suite de caractères de longueur variable prise dans un alphabet de 20 caractères qui représentent des acides aminés. L'ensemble E présente l'alphabet (ensemble des acides aminés). E={ a, c, d, e, f, g, h, i , k, l, m, n, p, q, r, s, t, v, w, y}. Le classement de protéines consiste à affecter automatiquement des nouvelles séquences à sa famille d'appartenance, c'est-à-dire dans la famille qui possède des séquences très similaires en structure et/ou en fonction Gibas et Jambeck (2002).

Dans cet article nous nous attachons à résoudre le problème de discrimination entre deux familles de protéines. Le travail est bien cadré, nous disposons d'un fichier d'apprentissage composé de deux familles de protéines à partir duquel nous construisons, à l'aide des techniques supervisées, une fonction de classement qui permet d'affecter une nouvelle séquence à une ou l'autre famille. Le problème du classement global, c'est-à-dire l'affectation d'une séquence à une famille parmi les innombrables catégories existantes est un peu plus ardue, notamment parce qu'il n'est pas réaliste de vouloir disposer d'un fichier d'apprentissage contenant toutes les familles existantes de protéines pour construire la fonction de classement, il apparaît clairement dans ce cas que les techniques supervisées classiques ne sont pas adaptées. Nous travaillons sur ce sujet actuellement.

Ce papier est organisé comme suit. Dans la section 2, on traite le problème de présentation des données biologiques et en particulier les protéines. Dans la section 3, on présente notre

processus de classement de protéines. La section 4 présente l'approche n-grammes d'extraction de descripteur. Dans la section 5, on discute le problème de pondération des descripteurs pour construire les tableaux de données. La phase de sélection de variables est détaillée dans la section 6. Enfin nous concluons notre article.

2 Problématique de représentation des séquences biologiques

Comme nous l'avons précisé, notre objectif est de mettre en place un système de classement qui, à partir de la description d'une séquence de protéine, lui affecte automatiquement sa famille d'appartenance. Pour ce faire, nous utilisons des données préalablement classées pour apprendre le système de classement avec des méthodes de fouille de données. Or, l'utilisation des données sous leur forme native n'est pas possible, les méthodes de fouille de données ne savent pas appréhender les séquences biologiques sous leur forme primaire, voir figure 2.

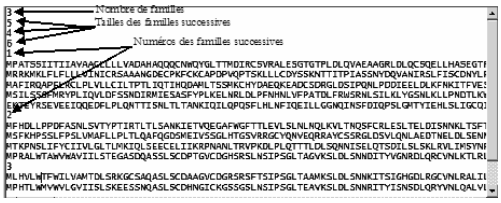

FIG. 2 – *Fichier de familles de protéines*

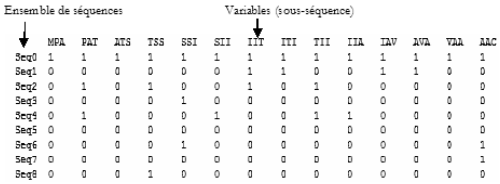

FIG. 3 – *Tableau individus/valeurs d'apprentissage*

Il est donc nécessaire de transformer ces données en tableau individus/variables avec, en ligne les séquences de protéines à analyser, en colonne les descripteurs extraits des protéines à analyser, et à l'intersection ligne-colonne, la valeur (la pondération) indiquant l'importance du descripteur dans la séquence. La figure 3 illustre un exemple de tableau booléen construit après l'extraction des variables.

A partir de cette représentation nous constatons la pertinence de l'analogie entre la représentation des séquences et celle des données textuelles. En effet, une séquence de protéines est une suite d'acides aminés, il y a 20 acides aminés possibles : une protéine est donc décrite par une chaîne de caractères représentant des acides aminés. Nous montrons dans la figure 2, un exemple de fichier décrivant quelques protéines. Cependant, a la différence du textmining, il n'existe pas de séparation naturelle dans les séquences de caractères, il n'est donc pas possible de mettre en exergue des " mots " auxquelles nous pourrions rattacher aisément une sémantique. Nous nous sommes donc tournés vers les techniques d'extraction de n-grammes, des suites de caractères, pour produire les descripteurs discriminants.

3 Processus de classement de protéine

Le processus respecte les étapes de l'ECD, il est composé de plusieurs étapes : Le rassemblement des familles de protéines où chaque famille est composée d'un certain nombre de séquences. Ces familles de protéines sont extraites à partir des banques de données biologiques

plus précisément des banques de protéines tels que SwissProt Gibas et Jambeck (2002), SCOP Murzin et al. (1995), etc.

- Le nettoyage des séquences consiste à éliminer les séquences inconnues et les séquences redondantes, dans ces deux étapes l'intervention d'un expert du domaine est obligatoire.
- L'extraction des variables prédictives à partir des données initiales. Ces variables sont appelées aussi des descripteurs discriminants.
- La sélection de variables, cette tâche consiste à identifier le meilleur sous-ensemble de descripteurs discriminants.
- L'apprentissage : elle consiste à apprendre la fonction de classement qui permet d'affecter automatiquement une protéine à sa famille d'appartenance.

Les différentes étapes de notre processus de classement sont présentées dans la figure 4.

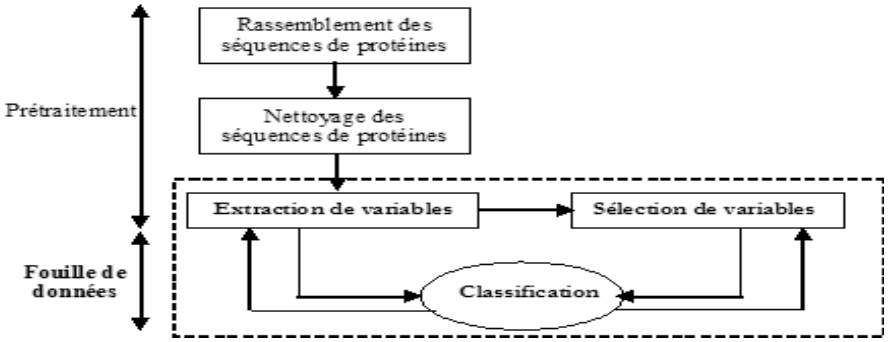

FIG. 4 – *Processus de classification de protéines*

L'extraction et la sélection des variables sont les phases les plus importantes dans notre travail. Dans la suite de l'article nous détaillerons ces deux tâches.

4 Extraction des descripteurs discriminants

Compte tenu de la nature des données que nous avons à manipuler, il semblait naturel d'utiliser des suites de caractères comme descripteurs discriminants. Ce choix est consécutif au rapprochement que nous avons réalisé avec la catégorisation de textes dans laquelle la technique des " n-grammes " a donné satisfaction dans de nombreux problèmes telles que la détection de courrier électroniques non-désirables Sebastiani (2002), la catégorisation de nouvelles Radwan et al. (2003). L'idée est donc de détecter les suites de caractères qui permettent de distinguer les différentes familles de protéines.

4.1 Les n-grammes

En général, un n-gramme est une suite de n caractères Miller et al. (1999). Pour un texte quelconque, l'ensemble des n-grammes qu'on peut générer est le résultat qu'on obtient en déplaçant une fenêtre de n case sur le corps du texte. Ce déplacement se fait par étapes, une étape (qui correspond à un caractère) pour chaque déplacement. A chaque étape une prise

de photo se fait, l'ensemble des photos constitue l'ensemble de tous les n-grammes qu'on puisse générer Miller et al. (1999). Les n-grammes qui constituent un outil efficace pour la catégorisation de textes, seront adaptés dans la classification de protéines.

Dans cet article, nous nous contentons d'extraire des descripteurs (n-grammes) de longueur fixe essentiellement dans le but de réduire les calculs, en plus à partir de ces descripteurs, on peut obtenir de bons résultats. Cependant, chercher des descripteurs de longueur variable est certainement une piste intéressante.

4.2 Identification de la meilleure taille des n-grammes(n)

Après l'identification de l'approche d'extraction, il reste à identifier la meilleure taille des n-grammes (n). Pour cela on a effectué plusieurs expérimentations sur des valeurs de n comprises entre 2 et 8. La table 2 de variation des n-grammes présente un très grand nombre de variables. A partir de la taille quatre des n-grammes on a trouvé des difficultés pour réaliser le classement : 31595 est un énorme nombre de n-grammes pour être traiter par les techniques de fouille de données. Notre choix de la valeur de n, commence à converger vers une valeur inférieure ou égale à 3.

A ce stade on ne peut pas favoriser une valeur sur les autres. Cependant on a appliqué une phase de sélection de variables. Cette phase est un enjeu fort, il y a des approches concurrentes. On peut présenter la sélection qui se base sur la fréquence, d'autres sur la corrélation entre variables, l'approche wrapper, etc. dans notre cas on a utilisé l'approche de sélection de variables qui se base sur la fréquence des n-grammes.
Notre but est d'identifier les n-grammes de faibles fréquences et de les ignorer par la suite. Cette stratégie consiste à éliminer les variables qui ont un pourcentage de présence inférieur ou égal à un taux (soit x%), où x est une constante qu'on a fait varier entre 5 et 25 voir tableau 1. La meilleure valeur de x sera conservée. Le tableau 1 montre que les fréquences des

N	Tous les N-grammes	Distinctes N-grammes	Filter <=05%	Filter <=15%	Filter<=25%
1	249	20	20	20	20
2	24962	400	400	396	387
3	59197	7145	4474	1287	464
4	65367	31595	2390	108	15
5	66125	42040	1344	25	4
6	66189	45206	1151	5	1
7	66135	46954	1045	1	0
8	66136	48226	984	0	0

TAB. 1 – *Filtrage des n-grammes.*

n-grammes de taille six, sept et huit sont trop faibles (<=5%). Et que ceux des n-grammes de taille quatre et cinq sont inférieurs ou égales à 25% sachant qu'ils sont trop nombreux (31595 pour les 4-grammes et 42040 pour les 5-grammes) donc on peut constater que ces n-grammes sont trop dispersés. Nous avons appliqué une deuxième approche complémentaire à celle de filtrage non supervisé. C'est une approche supervisée, elle consiste à évaluer l'ensemble de n-grammes par leurs taux de classification. Pour cela nous avons utilisé une representation

booléenne pour les tableaux des données, et comme classifieur nous avons appliqué le KNN. Les résultats fournis dans le tableau 2 montrent que les valeurs 2 et 3 de n sont les meilleures. On a conclu que les n-grammes de taille 3 proposent un bon compromis temps de calcul et

n	Distinct	TEr (1NN)
1	20	0.540230
2	400	0.034483
3	7145	0.057471
4	31595	0.103448
5	42040	0.183908
6	45206	0.252874
7	46954	0.275862
8	48226	0.298851

TAB. 2 – *Taux d'erreurs pour les différentes tailles de n (1-8)* .

performances en classement. Ce choix confirme les résultats décrits dans le domaine de la catégorisation de textes Sebastiani (2002). Ce travail à été publier dans Mhamdi et al. (2004).

5 pondérations des descripteurs

5.1 Présentation

Les techniques standard de fouille de données nécessitent un tableau de données. C'est un tableau d'attribut/valeur où chaque ligne représente une séquence de protéine et chaque colonne représente un 3-gramme. L'intersection entre une ligne et une colonne s'appelle la pondération des descripteurs. A la lumière des études réalisées dans la classification automatique de texte, nous avons testé 4 types de codages des données afin de déterminer quelle est la meilleure pondération dans notre problématique. Les pondérations testées sont les suivantes :

- Booléen : indique si un 3-gramme est présent ou non au sein d'une séquence ;
- Occurrence : nombre d'occurrences d'un 3-gramme dans une séquence ;
- Fréquence : fréquence relative d'un 3-gramme par rapport au nombre de 3-gramme composant une séquence ;
- TF*IDF : Corrige la fréquence de 3-gramme en fonction de sa fréquence au sein du fichier.

5.2 Expérimentations

Nos expérimentations ont été réalisées sur 5 familles de protéines extraites de la banque de données SCOP Gibas et Jambeck (2002). Nous évaluons les performances du type de représentation sur la discrimination deux à deux des familles de protéines. Les premiers résultats (tableau 4) montrent que les fichiers de fréquence et les fichiers d'occurrence semblent mieux adaptés et performants. On a utilisé le KNN comme classifieur et la méthode Leave-one-out pour évaluer les taux d'erreurs.

	Fich_Bool	Fich_Occ	Fich_FreqLong	Fich_TF_IDF
F_1_2	0.057471	0.045977	0.022989**	0.068966
F_1_3	0.053191	0.042553**	0.148936	0.063830
F_1_4	0.041322	0.033058**	0.132231	0.049587
F_1_5	0.046296	0.037037**	0.296246	0.064815
F_2_3	0.108911	0.099010	0.049505**	0.198020
F_2_4	0.062500**	0.085938	0.093750	0.218750
F_2_5	0.095652**	0.121739	0.260870	0.304348
F_3_4	0.177778	0.170370	0.096296**	0.229630
F_3_5	0.262295	0.278689	0.245902**	0.278689
F_4_5	0.161074	0.134228	0.040268**	0.100671

TAB. 3 – *Taux d'erreurs du classifieurs 1NN (** les meilleurs pour chaque test).*

On peut conclure que les matrices booléennes et d'occurrences sont les meilleures représentations. Et pour des raisons techniques on peut privilégié la représentation booléenne pour différentes raisons : stockage en mémoire des données ; elle convient autant aux techniques utilisant des variables continues (K-NN par ex.) que celles utilisant des variables discrètes (Naïve de Bayes par ex.) Mhamdi et al..

6 Classement et sélection de variables

Dans la phase précédente on s'est intéressé à l'extraction de variables. On a constaté que si on considère toutes les variables extraites, les performances des classifieurs sont dégradées. Cette dégradation peut être due à plusieurs raisons, telles que :

1. Elles sont produites automatiquement, certaines ne sont absolument pas utilisables pour le classement, d'autres ne sont pas pertinentes pour les familles que nous étudions.
2. Beaucoup d'entre elles sont redondantes, fortement corrélées, des méthodes d'apprentissage peuvent se révéler particulièrement inappropriées dans ce contexte.
3. Et enfin, la forte dimensionnalité (plusieurs milliers), surtout rapportée au faible effectif des séquences (une centaine) entraîne d'autre soucis pour l'apprentissage : temps de calcul prohibitifs, sur-apprentissage.

Alors comme solution on a opté pour une sélection de variables avant de faire la classification. La classification donc, sera réalisée avec les meilleures variables.

6.1 Les approches de sélections de variables

Dans la littérature il y a deux grandes familles de sélections de variables, la famille des méthodes de filtrages et la famille des méthodes wrapper.

1. L'approche filtre, on utilise des critères ad hoc pour sélectionner le meilleur ensemble de descripteurs, sans tenir compte des performances en apprentissage. Principal avantage : rapidité ; inconvénient : il est illusoire de prétendre trouver un ensemble de descripteurs qui serait le meilleur quelle que soit la méthode d'apprentissage.

2. L'approche wrapper qui optimise le taux d'erreur en utilisant explicitement la méthode d'apprentissage. Avantage : l'ensemble trouvé est optimal pour le classifieur étudié ; inconvénient : sur-apprentissage, estimation non biaisée de l'erreur, lenteur des calculs. Cette approche est possible mais n'est pas du tout réaliste dans notre contexte. Vue la grande masse de variables (descripteurs).

Dans cet article nous proposons une approche hybride. Dans cette approche nous essayons de combiner les avantages des deux approches précédentes. L'idée est de faire une présélection avec une approche filtre, et d'utiliser cette présélection pour déterminer la meilleure solution en évaluant le taux d'erreur d'une méthode d'apprentissage Sebban et Venturini (1999). Dans notre approche nous utilisons tout d'abord un critère de classement pour classer les variables après on applique un classifieur pour sélectionner les meilleures variables. Notre critère de classement se base sur la corrélation entre les variables.

6.2 Mesure de corrélation

Un descripteur X, prenant ses valeurs dans {x1,...,xL} est pertinent pour la prédiction des valeurs de l'attribut classe C {c1,...,cK} si, pour un individu donné, à une valeur de X on peut associer une et une seule valeur de C. En d'autres termes, la connaissance que l'on a de X permet de réduire l'incertitude quant aux valeurs prises par C. Parmi les nombreux indicateurs qui permettent de traduire degré de dépendance entre deux variables, nous avons choisi les mesures de corrélation. Si la corrélation entre deux attributs continus est bien connue depuis très longtemps, il existe en revanche plusieurs interprétations en ce qui concerne la corrélation entre deux attributs discrets. Dans cet article nous choisirons une interprétation en termes d'information mutuelle. La mesure est normalisée afin que le cœfficient de corrélation qui en est déduit varie entre 0 (absence de lien entre Y et X) et 1 (liaison fonctionnelle). Cette mesure est symétrique, elle possède la propriété très importante d'être insensible au nombre de modalités, permettant ainsi de comparer différents descripteurs candidats. Il existe plusieurs critères de mesure de corrélation, tels que $\chi 2$,ρ, etc. Hastie et Friedman (2001)

6.3 Combiner le classement de variables et l'évaluation wrapper (SHFR)

6.3.1 Principe

Le principal écueil de la sélection de variable fondée sur la corrélation est le choix de la règle d'arrêt. Les descripteurs étant évalués indépendamment de la méthode d'apprentissage, il est très difficile de déterminer la taille optimale du sous-ensemble de descripteurs sélectionnés. L'approche hybride permet de dépasser cet écueil. Elle propose de tirer parti de la rapidité de constitution des solutions successives par le calcul de la corrélation, tout en utilisant par la suite une évaluation des performances des solutions avec une méthode de ré-échantillonnage. Une première définition de la méthode hybride serait dès lors (a) de calculer pour chaque attribut candidat sa corrélation, (b) s'appuyer sur le classement obtenu pour les trier, (c) tester les sous-ensembles d'attributs de taille croissante en calculant le taux d'erreur en validation croisée d'un algorithme d'apprentissage donné en prenant tout d'abord le premier, puis les deux premiers, etc. (d) jusqu'à ce qu'on ait testé tous les sous-ensembles possibles. Notons que les solutions évaluées sont imbriquées : la solution à l'étape j contient les (j-1) premiers attributs plus la jème ; le nombre de fois où l'on fera appel à la méthode d'apprentissage est

connu à l'avance, il est égal au nombre de descripteurs dans la base, nous verrons plus loin qu'il est possible également d'introduire comme paramètre le nombre maximal de solutions à tester Duch et al. (2004). Dans notre expérimentation, nous avons utilisé le t de Tschuprow pour classer les variables, il s'agit d'une normalisation du CHI-2, dans le cadre de variables binaires, il peut s'interpréter comme une corrélation Mhamdi et al. (2005), ce travail est intitulé SHFR (Standard Hybrid Feature Ranking).

6.3.2 Expérimentations

On a réalisé des expérimentations sur ce point. Les résultats sont montrés dans le tableau 4 et la figure 5.

Familles	Nombre de 3-grammes	Taux d'erreur après la selection de 3-grammes
F_1_2	11	0
F_1_3	12	0
F_1_4	5	0
F_1_5	2	0
F_2_3	30	0.019802
F_2_4	62	0.007813
F_2_5	83	0.026087
F_3_4	13	0.022222
F_3_5	16	0.040984
F_4_5	9	0.026846

TAB. 4 – *Taux d'erreur avant et après la sélection de variable .*

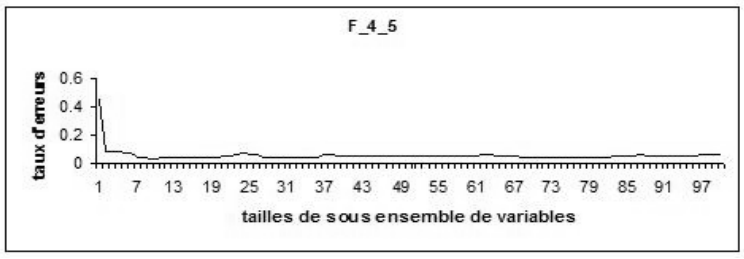

FIG. 5 – *variation du Taux d'erreurs de classification avec le nombre des n-grammes*

6.4 Prise en considération des redondances (RBHFR)

6.4.1 Principe

Cette première approche que nous avons par ailleurs expérimentée dans nos travaux antérieurs donne d'excellents résultats. Elle présente néanmoins une faille importante, elle ne permet pas de tenir compte de la redondance des attributs. Ainsi, si l'on duplique 9 fois l'attribut le plus corrélé avec la classe dans la base, les dix premières places seront occupées par le même attribut. Il est donc nécessaire de nous tourner vers des mesures d'évaluation qui, pour chaque descripteur candidat, tient à la fois compte de sa corrélation avec l'attribut classe variable mais aussi de son degré de dépendance avec les attributs déjà insérés dans le classement. Dans cet article, nous nous sommes tournés vers les techniques de sélection de variables qui ont justement traité de la corrélation entre les descripteurs. Au premier passage, la première variable sélectionnée sera toujours celle qui est la plus corrélée avec l'attribut classe, mais au passage suivant, nous classerons en second l'attribut qui sera à la fois la plus corrélée avec l'attribut classe et la moins corrélée avec les attributs déjà sélectionnés. Pour mesurer ce compromis, nous avons utilisé la formule proposée dans l'algorithme CFS Hall (2000).

$$SU[Y,X] = 2 * [\frac{H(Y) - H(Y/X)}{H(X) + H(Y)}] \quad (1)$$

Cependant, à la différence de ces techniques qui cherchent à définir de manière ad hoc, indépendamment de la méthode d'apprentissage, la taille adéquate du meilleur sous ensemble de descripteurs, nous introduisons en paramètres le nombre d'attributs maximum à introduire. Schématiquement, nous avons donc supprimé la règle d'arrêt de l'algorithme originel et nous appliquons la méthode d'apprentissage pour détecter le meilleur sous-ensemble de descripteurs en nous appuyant sur l'ordre établi par la sélection. Notre espoir est qu'en éliminant autant que possible la redondance des attributs dans les solutions successives présentés à la méthode d'apprentissage, le sous-ensemble d'attributs sélectionné au final sera de taille significativement réduite par rapport à une approche de classement de variable classique. Dans un deuxième travail intitulé RBHFR (Redundancy Based Hybrid Feature Ranking) Rakotomalala et al. (2005), nous avons voulu améliorer les résultats du premier travail, en effet la première méthode présente néanmoins une faille importante, elle ne permet pas de tenir compte de la redondance des attributs. Pour cela, nous nous sommes tournés vers les techniques de sélection de variables fondées sur les corrélations qui tiennent compte des inter-corrélations entre les descripteurs. Pour mesurer ce compromis, nous avons utilisé la formule proposée dans l'algorithme CFS Hall (2000).

6.4.2 Expérimentations

	Meilleur erreur		Nombre de 3-grammes			Erreur moyenne	
Protein pair	SHFR	RBHFR	SHFR	RBHFR	CFS	SHFR	RBHFR
F12	0.00	0.00	5	5	29	0.038	0.013
F13	0.00	0.00	5	5	16	0.038	0.021
F14	0.016	0.00	8	3	27	0.025	0.013
F15	0.009	0.009	3	3	18	0.029	0.021
F23	0.039	0.00	13	7	56	0.103	0.032
F24	0.031	0.008	6	20	32	0.053	0.031
F25	0.052	0.017	12	8	39	0.096	0.042
F34	0.021	0.021	4	4	12	0.032	0.029
F35	0.057	0.031	16	12	28	0.095	0.052
F45	0.033	0.020	8	12	6	0.046	0.036

TAB. 5 – *Comparaison entre les deux méthodes SHFR et RBHFR.*

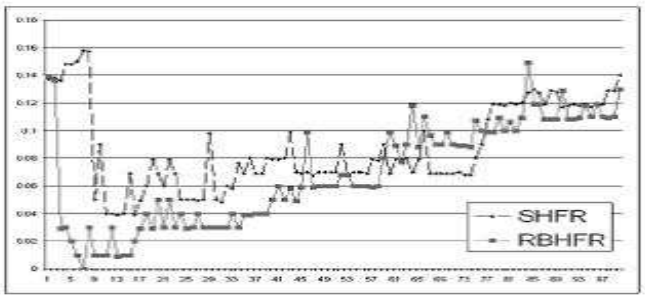

FIG. 6 – *Comparaison entre les deux approches hybrides de sélection de n-grammes*

Les résultats présentés par le tableau 5 et la figure 6 montrent que RBHFR fournis des taux d'erreurs inférieurs à ceux de l'approche SHFR.

7 conclusion

Dans cet article, nous avons tracé les grandes lignes du processus de classement automatique de protéines à partir de leurs structures primaires en nous appuyant sur le cadre générique de l'ECD, plus précisément en exploitant l'analogie qui existe entre notre problématique et celle du classement automatique de textes. Les techniques que nous avons développées ont été

évaluées sur des séquences réelles de protéines issues de banques de données accessibles sur internet.

Le processus de classement passe par plusieurs étapes. Dans un premier temps, il s'agit d'extraire des descripteurs pertinents à partir de la représentation primaire et proposer une pondération qui permet de produire des fonctions de classement performants : nous avons montré que les 3-grammes et la pondération booléenne (présence/absence) donne des résultats très satisfaisants. Dans un deuxième temps, nous nous sommes intéressés à la réduction de l'espace de représentation en essayant de mettre en avant les descripteurs les plus performants. Au vu du nombre important de descripteurs candidats, chercher le sous-ensemble optimal à l'aide de la méthode " wrapper ", c'est-à-dire minimisant le taux d'erreur du classifieur utilisé, n'est pas réalisable. Il s'agissait donc de réduire l'espace de recherche en utilisant les méthodes de filtrage, l'idée est de produire des sous-ensembles imbriqués de descripteurs candidats, d'exploiter cet ordonnancement pour rechercher celui qui minimise l'erreur de prédiction. Nos expérimentations ont montré que l'on combinait ainsi les avantages des deux méthodes : rapidité d'exploration, et efficacité des solutions produites.

Notre approche est restreinte sur les problèmes de discrimination entre deux familles de protéines. Notre prochain enjeu sera de proposer une approche qui permettra de reconnaître une famille particulière de protéine parmi les autres, en sachant que dans l'apprentissage aura été effectué sur un univers incomplet, à savoir que certaines familles n'auront pas été présentes dans le fichier d'apprentissage. Nos premières évaluations montrent que les techniques classiques d'apprentissage supervisé sont inopérantes dans ce contexte, il nous faut donc redéfinir notre démarche et évaluer des techniques plus appropriées pour résoudre ce problème.

Tous nos travaux sont suivis et validés par des biologistes de l'Institut Pasteur de Tunis.

Références

Duch, W., T. Wieczorek, J. Biesiada, et M. Blachnik (2004). Comparison of feature ranking methods based on information entropy. *Proc. of International Joint Conference on Neural Networks (IJCNN, IEEE Press*, 1415–1420.

Fayyad, U. M., G. Piatetsky-Shapiro, et P. Smyth (1996). From data mining to knowledge discovery: An overview, in " advances in knowledge discovery and data mining. *AAAI Press and the MIT Press*, 1–34.

Gibas, C. et P. Jambeck (2002). *Introduction à la bioinformatique*. Oreilly.

Hall, M. (2000). *Correlation-based feature selection for discrete and numeric class machine learning*. Berlin: In ICML'00, proceedings of the 17th International conference on machine learning,Morgan kaufman publishers Inc.

Hastie, T.and Tibshirani, R. et J. Friedman (2001). The elements of statistical learning. *Springer-Verlag*.

Mhamdi, F., M. Elloumi, et R. Rakotomalala. *in Proc. Neuro-Computing and Evolving Intelligence, NCEI'2004*.

Mhamdi, F., M. Elloumi, et R. Rakotomalala (2004). *Textmining, features selection and datamining for proteins classification*. Proc. of ICTTA'04 International Conference on In-

formatique & Communication technologies :From Theory to Application, ICTTA'04, IEEE (Damascus, Syria), IEEE Catalog Number : 04EX852C.

Mhamdi, F., R. Rakotomalala, et M. Elloumi (2005). *Feature Ranking for Protein Classification*. in Proc. 4th International Conference on Computer Recognition Systems, Springer-Verlag.

Miller, D., D. J. Shen Liu, et C. Nicholas (1999). Performance and scalability of a large-scale n-gramme based information retrivial system. *Journal of digital information 1(5)*.

Murzin, G. A., E. S. Brenner, T. Hubbard, et C. Chothia (1995). Scop: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Bio.. 247*, 536–540.

Radwan, J., C. Jérémy, et R. Rakotomalala (2003). Un cadre pour la catégorisation de textes multilingues. *In Proc of 7èmes Journées internationales d'Analyse statistique des Données Textuelles ...*, 650–660.

Rakotomalala, R., F. Mhamdi, et M. Elloumi (2005). *Hybrid Feature Ranking for Protein Classification*. Berlin: in Proc. 1st International Conference on Advanced Data Mining and Applications, Springer Lecture Notes in Computer Science series, Springer-Verlag.

Sebastiani, F. (2002). Machine learning in automated text categorisation. *ACM computing Surveys 34*, 1–47.

Sebban, M. et G. Venturini (1999). *Apprentissage automatique*. Hermes Sciences.

Summary

The classification of proteins is an important activity for the biologist. With the increasing increase of the size of data banks, the manual classification became impossible, it became so necessary to implement computer strategies to automate process. We present a global frame inspired of the walk of the knowledge descovery in database to classify automatically proteins from their primary structures. This frame contains two great stages: the first, the phase of meadow-treatment, consists in extracting descriptors from the original data description, we used the very known technique of n-grams in the categorization of texts; second, the phase of learning, consists in using the techniques of learning, integrating the phase of feature selection, to classify automatically proteins. Experiments on real data banks allow to obtain encouraging results. This article reviews the various alternatives that we analyzed in every stage of the classification process.