TALN 2009, Senlis, 24–26 juin 2009

Motifs séquentiels pour l'extraction d'information : illustration sur le problème de la détection d'interactions entre gènes

Marc Plantevit Thierry Charnois
GREYC – CNRS UMR 6072
Université de Caen – Bd Mal. Juin
14032 CAEN Cedex
prénom.nom@info.unicaen.fr

Résumé. Face à la prolifération des publications en biologie et médecine (plus de 18 millions de publications actuellement recensées dans PubMed), l'extraction d'information automatique est devenue un enjeu crucial. Il existe de nombreux travaux dans le domaine du traitement de la langue appliquée à la biomédecine ("BioNLP"). Ces travaux se distribuent en deux grandes tendances. La première est fondée sur les méthodes d'apprentissage automatique de type numérique qui donnent de bons résultats mais ont un fonctionnement de type "boite noire". La deuxième tendance est celle du TALN à base d'analyses (lexicales, syntaxiques, voire sémantiques ou discursives) coûteuses en temps de développement des ressources nécessaires (lexiques, grammaires, etc.). Nous proposons dans cet article une approche basée sur la découverte de motifs séquentiels pour apprendre automatiquement les ressources linguistiques, en l'occurrence les patrons linguistiques qui permettent l'extraction de l'information dans les textes. Plusieurs aspects méritent d'être soulignés : cette approche permet de s'affranchir de l'analyse syntaxique de la phrase, elle ne nécessite pas de ressources en dehors du corpus d'apprentissage et elle ne demande que très peu d'intervention manuelle. Nous illustrons l'approche sur le problème de la détection d'interactions entre gènes et donnons les résultats obtenus sur des corpus biologiques qui montrent l'intérêt de ce type d'approche.

Abstract. The proliferation of publications in biology and medicine (more than 18 million publications currently listed in PubMed) has lead to the crucial need of automatic information extraction. There are many work in the field of natural language processing applied to biomedicine (BioNLP). Two types of approaches tackle this problem. On the one hand, machine learning based approaches give good results but run as a "black box". On the second hand, NLP based approaches are highly time consuming for developing the resources (lexicons, grammars, etc.). In this paper, we propose an approach based on sequential pattern mining to automatically discover linguistic patterns that allow the information extraction in texts. This approach allows to overcome sentence parsing and it does not require resources outside the training data set. We illustrate the approach on the problem of detecting interactions between genes and give the results obtained on biological corpora that show the relevance of this type of approach.

Mots-clés : Extraction d'information, fouille de textes, motifs séquentiels, interactions entre gènes.

Keywords: Information extraction, text mining, sequential patterns, gene interactions.

1 Introduction

Le volume des publications dans le domaine de la biologie et de la médecine s'accroit à un rythme considérable : plus de 16 millions de publications recensées dans la base MedLine et disponibles via PubMed[1] en 2005 et une augmentation quotidienne de 1800 références (chiffres de 2005 tirés du site de l'INSERM). Dans cette masse de données textuelles, la recherche manuelle d'information est impossible. Deux types de requêtes intéressent le biologiste : "dans quels articles parle-t-on du gène X ?", "Avec quel(s) gène(s), le gène X interagit-il ?, et sous quelle forme ?".

Depuis une bonne quinzaine d'années de nombreux travaux en extraction d'information et en fouille de textes appliquées au domaine biomédical ont vu le jour. Deux tâches sont particulièrement explorées correspondant aux deux requêtes mentionnées précédemment : la première est la reconnaissance d'entités nommées de type biologique (noms de gènes, protéines, fonctions biologiques, etc.) et la deuxième concerne l'identification et le typage de relations entre entités biologiques précédemment reconnues. Dans cet article, nous nous intéressons à la deuxième tâche et plus particulièrement à celle consistant à détecter les interactions entre gènes et à leur typage (inhibition, simulation,etc.) et à terme à ce que nous appelons « modalité » et qui peut intervenir dans la relation (négation, interaction possible, démontrée, etc.).

Deux types d'approches sont généralement considérés (Zweigenbaum *et al.*, 2007). Le premier est fondé sur des méthodes statistiques ou d'apprentissage : les approches à base de calcul de co-occurrences sont simples à mettre en oeuvre et peuvent donner un bon rappel, mais les résultats sont faibles en précision (deux entités peuvent apparaître dans une énumération sans interagir), et ne donnent aucune caractérisation de l'interaction. Les autres approches en apprentissage se ramènent à des problèmes de classification comme spécifié dans le challenge BioCreative II (Krallinger *et al.*, 2008) : par exemple, décider à partir du résumé si l'article est susceptible de contenir une interaction. Pour ce type de tâche, les meilleurs résultats sont obtenus par des machines à vecteurs de support (SVM), ou encore les « conditional random fields » (CRFs) qui utilisent des descripteurs statistiques. Ces approches ont un fonctionnement de type « boite noire » (les règles ne sont ni interprétables ni modifiables par un expert), la mise en place des descripteurs est lourde à mettre en œuvre, et ne donnent pas le type de l'interaction.

A l'opposé, les approches de type TAL s'appuient sur des connaissances linguistiques : règles d'extraction, analyse syntaxique voire sémantique de la phrase, pour découvrir les interactions (Zweigenbaum *et al.*, 2007). Une règle d'extraction, ou patron linguistique, peut s'exprimer sous forme d'expression régulière à base de mots ou d'étiquettes grammaticales et d'un verbe, ou d'une nominalisation d'un verbe, exprimant une interaction (interact, bind, encode). Généralement, l'application des patrons s'opère après l'analyse de la phrase pour améliorer les résultats. En dépit des progrès récents effectués dans le domaine de l'analyse syntaxique, l'extraction reste fortement dépendante des résultats de la syntaxe. Par ailleurs, ce type de méthodes a un coût important en termes d'écriture et de développement des patrons. Pour contourner ce problème, certaines approches visent à apprendre automatiquement les règles d'extraction, mais l'apprentissage est réalisé sur l'analyse syntaxique des phrases comme dans (Vetah *et al.*, 2004; Kim *et al.*, 2007).

Nous proposons dans cet article une approche pour apprendre les patrons linguistiques par une méthode de fouille de données basée sur les motifs séquentiels. A notre connaissance, les mo-

[1] http://www.ncbi.nlm.nih.gov/pubmed/

tifs séquentiels n'ont pas encore été utilisés pour réaliser de l'extraction d'information dans des textes biomédicaux. L'originalité principale de l'approche consiste à s'émanciper de l'analyse syntaxique pour l'apprentissage des patrons et pour leur application. Elle ne nécessite pas de ressources linguistiques autres que le corpus d'apprentissage. Celui-ci consiste en un ensemble de phrases qui contiennent des interactions et dont les entités nommées de type gène sont annotées en tant que tel. Seul un pré-traitement est réalisé sur le corpus : tous les mots sont remplacés par leur lemme auquel est ajouté la catégorie morpho-syntaxique[2] du mot. De la même manière, l'application des patrons pour extraire les interactions entre gènes s'effectue de façon directe et sans utiliser d'analyse syntaxique.

Nous décrivons la méthode dans la section 2 en détaillant les différentes phases : extraction des motifs séquentiels fréquents, ajout de contraintes et sélection. Puis nous présentons l'expérimentation menée et en discutons les résultats (section 3) qui montrent l'intérêt de l'approche.

2 Méthode

2.1 Aperçu de la méthode

La figure 1 décrit le fonctionnement général de notre méthode. Tout d'abord, une base de séquences textuelles est créée à partir d'un ensemble de phrases contenant des interactions entre gènes sur lesquelles une analyse morpho-syntaxique a été effectuée (étape 1). Dans l'étape 2, les motifs séquentiels fréquents sont extraits à partir de la base de séquences textuelles. L'ensemble des motifs séquentiels fréquents peut être relativement important, rendant impossible toute utilisation future des motifs découverts. L'étape 3 vise à contraindre ces motifs afin de sélectionner un sous ensemble de motifs séquentiels qui respectent un ensemble de contraintes. Les motifs satisfaisant les contraintes sont alors divisés en plusieurs sous ensembles par rapport aux verbes et aux noms qu'ils contiennent. Par exemple, un ensemble contiendra tous les motifs séquentiels contenant l'élément *interact@vvz*. Etant donné un entier k fixé *a priori*, chaque sous-ensemble est alors fouillé récursivement afin de disposer d'au plus k représentants par sous-ensemble. L'étape 4 est dédiée à cette tâche. Le nombre de motifs séquentiels restants peut être alors facilement examiné par un expert humain. L'étape 5 représente la validation des motifs par un expert. Les motifs séquentiels validés forment alors l'ensemble des patrons linguistiques qui sont ensuite utilisés pour détecter des interactions entre gènes dans des textes biomédicaux (étape 6).

Nous décrivons plus précisément l'ensemble de ces étapes dans la suite de cet article.

2.2 Extraction de motifs séquentiels fréquents

Introduite par (Srikant & Agrawal, 1996), l'extraction de motifs séquentiels fréquents permet de découvrir des corrélations entre des événements au cours d'une relation d'ordre (e.g., le temps). Ce problème est devenu au fil des années un domaine actif de la fouille de données avec de nombreux algorithmes à la clé (Pei *et al.*, 2001; Zaki, 2001).

[2] Dans cet article, nous appelons morpho-syntaxique, l'analyse au niveau du mot fournissant sa catégorie grammaticale, parfois associée à une information morphologique comme le mode passif pour un verbe.

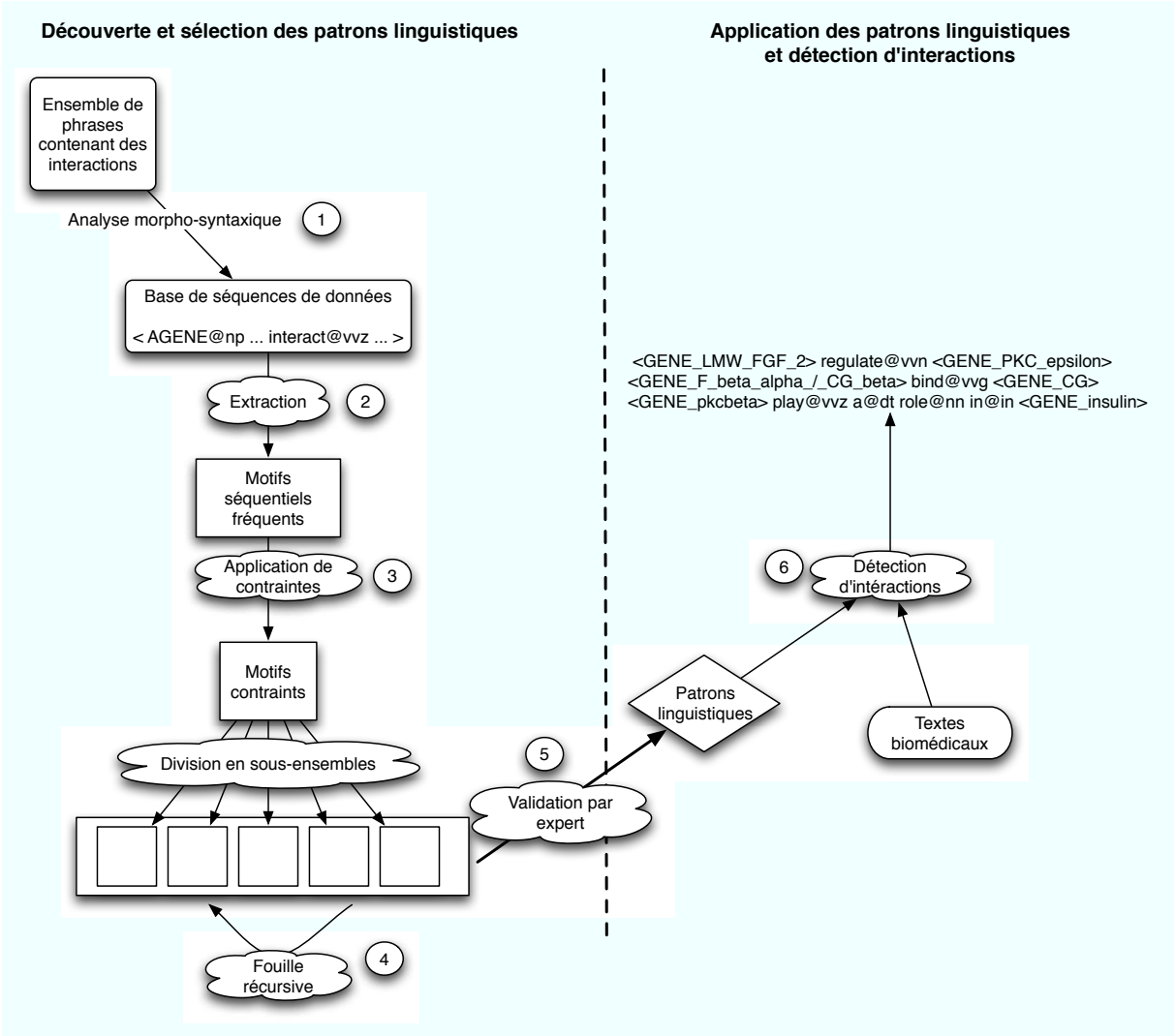

FIG. 1 – Schéma général de l'approche

Etant donné un ensemble $\mathcal{I}$ de littéraux distincts appelés *items*, une séquence $s = \langle i_1, i_2, \ldots, i_n \rangle$ est une liste ordonnée non vide d'items. Une séquence $S_a = \langle a_1, a_2, \ldots, a_n \rangle$ est incluse dans une autre séquence $S_b = \langle b_1, b_2, \ldots, b_m \rangle$ s'il existe des entiers $1 \leq i_1 < i_2 < \ldots < i_n \leq m$ tels que $a_1 = b_{i_1}, a_2 = b_{i_2}, \ldots, a_n = b_{i_n}$. Si la séquence S_a est incluse dans S_b, alors S_a est une sous-séquence de S_b et S_b est une super-séquence de S_a, noté $S_a \preceq S_b$. Une base de séquences SDB est un ensemble de paires (sid, S) où sid est un identifiant de séquence et S est une séquence. Une paire (sid, S) *contient* une séquence S_α si S_α est une sous-séquence de S ($S_\alpha \preceq S$). Le support absolu d'une séquence S_α dans une base de séquences SDB correspond au nombre de paires (sid, S) qui contiennent S_α. Le support relatif représente le pourcentage de paires qui supportent S_α ($\frac{|(sid,S)\ t.q.\ S_\alpha \preceq S|}{|SDB|}$).

Etant donnés une base de séquences SDB et un seuil de support minimum $supmin$, le problème de l'extraction de motifs séquentiels fréquents est de retourner l'ensemble complet FS des séquences S_α contenues dans SDB qui ont un support supérieur ou égal à $supmin$ ($support(S_\alpha) \geq supmin$).

2.3 Extraction de motifs séquentiels dans des données textuelles

Nous montrons ici comment nous appliquons la découverte de motifs séquentiels à des données textuelles. En particulier, nous décrivons le choix de la base de séquences SDB dans un contexte de données textuelles.

Nous disposons d'un ensemble de phrases reconnues comme contenant des interactions entre gènes. Les gènes sont déjà étiquetés comme tels dans ces phrases. Dans cet article, nous considérons les phrases contenant des interactions et au moins deux noms de gènes afin d'éviter le problème introduit par les structures anaphoriques qui reste un problème non résolu (Zweigenbaum *et al.*, 2007). Ces phrases sont issues d'articles accessibles sur PubMed. Notre objectif étant d'apprendre des motifs représentant des interactions entre gènes, nous considérons la phrase comme séquence de données. Les items sont les lemmes auxquels sont associés leur étiquette morpho-syntaxique et la relation d'ordre est l'ordre des mots dans la phrase.

Par exemple, une phrase contenant une interaction est : « *Here we show that <Gene SOX10>, in synergy with <GenePAX3>, strongly activates <Gene MITF > expression in transfection assays.* »

Nous remplaçons les noms de gènes par un item particulier noté *AGENE*. Ensuite, nous procédons à un étiquetage morpho-syntaxique des phrases. Les séquences de SDB seront alors les phrases étiquetées par un analyseur morpho-syntaxique.

Ainsi, la phrase précédente devient une séquence de SDB :

⟨ *here@rb we@pp show@vvp that@in/that AGENE@np ,@, in@in synergy@nn with@in AGENE@np ,@, strongly@rb activate@vvz AGENE@np expression@nn in@in transfection@nn assay@nns .@sent* ⟩

En fouille de motifs fréquents, le choix du seuil de support minimum est un problème récurrent. Si le seuil de support est trop élevé, le risque est d'extraire uniquement des généralités qui n'apporteront rien à l'utilisateur. Si le seuil de support est trop faible, l'ensemble des motifs fréquents extraits peut être extrêmement volumineux rendant impossible toute utilisation. Dans cet article, nous faisons le choix de considérer un seuil de support faible et de réduire l'ensemble des motifs séquentiels fréquents en post-traitement en introduisant des contraintes supplémentaires différentes de la contrainte de fréquence utilisée ($supmin$).

2.4 Ajout de contraintes

En fouille de motifs, les contraintes permettent à l'utilisateur de définir plus précisément ce qu'il considère comme intéressant. Ainsi, la contrainte la plus utilisée est la contrainte de fréquence ($supmin$) qui permet de considérer les motifs qui respectent cette contrainte. Il est possible d'utiliser différentes contraintes en plus de la fréquence (Ng *et al.*, 1998). Notons que dans cet article, nous appliquons les contraintes autres que la fréquence en post-traitement de l'ensemble des motifs fréquents.

Etant donné que nous souhaitons extraire des motifs séquentiels qui modélisent des interactions entre gènes, nous pouvons utiliser les contraintes suivantes :

– $\mathcal{C}_{2g}$ impose qu'un motif séquentiel doit contenir au moins deux fois l'item *AGENE*. Ainsi l'ensemble $SAT(\mathcal{C}_{2g})$ qui représente l'ensemble des motifs qui satisfont la $\mathcal{C}_{2g}$ est égal à : $SAT(\mathcal{C}_{2g}) = \{s = \langle i_1, \ldots, i_n \rangle \; t.q. \; |\{j \; t.q. \; i_j = \textit{AGENE}\}| \geq 2\}$.

– $\mathcal{C}_{vn}$ impose qu'un motif séquentiel doit contenir au moins un nom ou un verbe. L'ensemble $SAT(\mathcal{C}_{vn})$ des motifs séquentiels qui satisfont $\mathcal{C}_{vn}$ est :
$SAT(\mathcal{C}_{vn}) = \{s = \langle i_1, \ldots, i_n \rangle \; t.q. \; \exists i_j, \; verbe(i_j) = vrai \vee nom(i_j) = vrai\}$.
– Afin de réduire la redondance des motifs séquentiels, nous pouvons considérer les séquences fréquentes maximales (par rapport à l'inclusion $\preceq$). Un motif séquentiel fréquent s_1 est maximal s'il n'existe pas de motif séquentiel fréquent s_2 tel que $s_1 \preceq s_2$. Nous notons $\mathcal{C}_{max}$ cette contrainte. L'ensemble $SAT(\mathcal{C}_{max})$ des motifs séquentiels fréquents qui vérifient cette contrainte est : $SAT(\mathcal{C}_{max}) = \{s \; t.q. \; support(s) \geq supmin \wedge \nexists s' \; t.q. \; support(s') \geq supmin, \; s \preceq s'\}$.

Les contraintes précédentes peuvent être regroupées en une unique contrainte $\mathcal{C}_G$ qui est la conjonction de ces trois contraintes. L'ensemble $SAT(\mathcal{C}_G)$ des motifs séquentiels fréquents qui vérifient la contrainte $\mathcal{C}_G$ est égal à $SAT(\mathcal{C}_{2g}) \bigcap SAT(\mathcal{C}_{vn}) \bigcap SAT(\mathcal{C}_{max})$.

Même si l'ensemble $SAT(\mathcal{C}_G)$ est sensiblement plus petit que l'ensemble complet des motifs séquentiels fréquents, il est possible que cet ensemble soit encore trop important pour être analysé et validé par un utilisateur humain.

2.5 Extraction récursive de motifs séquentiels

Nous divisons l'ensemble $SAT(\mathcal{C}_G)$ en plusieurs sous-ensembles E_{X_i} où le sous-ensemble E_{X_i} regroupe tous les motifs séquentiels de $SAT(\mathcal{C}_G)$ contenant l'item X_i. Plus formellement, $E_{X_i} = \{s \in SAT(\mathcal{C}_G) \, t.q. \, \langle X_i \rangle \preceq s\}$. Notons que nous réduisons les X_i aux éléments étiquetés comme étant un verbe ou un nom.

Nous souhaitons alors déterminer au plus k ($k \geq 1$) représentants pour chaque ensemble E_{X_i}. Les principes de la fouille récursive introduite par (Soulet, 2007), qui visent à exhiber des représentants parmi des motifs ensemblistes émergents, s'appliquent dans notre contexte. Chaque sous-ensemble E_{X_i} est ainsi fouillé récursivement avec un seuil de support minimum $minsup$ égal à $\frac{1}{k}$ afin d'extraire les motifs séquentiels fréquents vérifiant la contrainte globale $\mathcal{C}_G$ introduite précédemment. La récursivité s'arrête dès que le nombre de motifs séquentiels extraits vérifiant $\mathcal{C}_G$ est inférieur ou égal à k[3]. En d'autres termes, dans la fouille récursive les motifs séquentiels extraits sur une base de séquences deviennent à leur tour la base de séquences qui va être fouillée. Ce processus se termine quand le nombre de motifs extraits sur une base de séquences (motifs) est inférieur ou égal à k.

Pour chaque sous-ensemble E_{X_i}, les k motifs séquentiels extraits récursivement sont des motifs séquentiels qui sont fréquents sur la base de séquence SDB. Autrement dit, ils appartiennent à l'ensemble complet des motifs séquentiels fréquents dans SDB par rapport à $supmin$.

A la fin de cette étape, le nombre de motifs séquentiels représentant des interactions entre gènes est maîtrisé. Il est inférieur ou égal à $n \times k$ où n est le nombre de sous-ensembles E_{X_i} de $SAT(\mathcal{C}_G)$. Notons que le paramètre k est fixé *a priori* par l'utilisateur. Ainsi, l'ensemble des motifs séquentiels représentant des interactions entre gènes peut être analysé par un utilisateur humain car sa taille le permet. Les motifs séquentiels sont alors validés par l'utilisateur et forment des patrons linguistiques permettant la détection d'interaction entre gènes. De plus, il est intéressant de noter que la sous-catégorisation du verbe donnée par l'étiquetage morpho-syntaxique indique la forme passive ou active du verbe et permet de repérer le sens de l'interaction. Les prépositions permettent aussi de repérer cette information lorsque le patron ne contient

[3]La contrainte $\mathcal{C}_{max}$ permet d'assurer la terminaison de la récursivité.

pas de verbe.

3 Expérimentation et résultats

Nous avons effectué des expérimentations de notre méthode. Dans cette section, nous présentons d'abord l'acquisition et la validation des patrons linguistiques et ensuite leur application sur des jeux de données réels.

3.1 Découverte des patrons linguistiques

Les gènes peuvent interagir entre eux par l'intermédiaire des protéines qu'ils synthétisent. De plus, bien qu'il existe des conventions, les biologistes ne font généralement pas de différence dans les textes entre le nom du gène et le nom de la protéine synthétisée par le gène. Ils écrivent l'un pour l'autre, et savent en fonction du contexte si la phrase traite de la protéine ou du gène. Ainsi, pour découvrir les patrons linguistiques d'interaction entre les gènes, nous avons réuni deux corpus différents contenant des noms de gènes et de protéines.

Le premier corpus contient des phrases issues de résumés de PubMed, annotées par Christine Brun de l'Institut de Biologie du Developpement de Marseille-Luminy. Il contient 1806 phrases annotées. Ce corpus est disponible en tant que source secondaire d'apprentissage de la tâche « Protein-protein Interaction task (Interaction Sentence Sub-task, ISS) » du challenge BioCreative II (Krallinger *et al.*, 2008).

Le second corpus contient des phrases d'interactions entre protéines annotées par un expert. Ce jeu de données qui contient 2995 phrases d'interaction est décrit dans (Rosario & Hearst, 2005).

Nous avons fusionné les deux corpus et attribué une étiquette unique aux différents noms de gènes et protéines : *AGENE*. Une analyse morpho-syntaxique est ensuite effectuée à l'aide de l'analyseur *TreeTagger* (Schmid, 1994). Les phrases sont alors prêtes pour être fouillées afin d'extraire l'ensemble des motifs séquentiels fréquents. Nous fixons un seuil de support minimum égal à 10. En effet, un tel seuil permet de ne pas considérer le bruit tout en permettant la découverte de nombreuses formes d'interactions. Le nombre de motifs séquentiels fréquents extraits est relativement important. Plus de 32 millions de séquences sont découvertes. Bien que le nombre de motifs extraits soit très important, l'extraction de cet ensemble a nécessité 15 minutes. L'extracteur utilisé est $dmt4$ développé par C. Rigotti (Nanni & Rigotti, 2007).

L'application des contraintes $\mathcal{C}_{2g}$, $\mathcal{C}_{vn}$ et $\mathcal{C}_{max}$ permet de réduire sensiblement le nombre de motifs séquentiels considérés. En effet, le nombre de motifs séquentiels satisfaisant les trois contraintes est d'environ 65 000. Toutefois, ce nombre reste prohibitif pour une analyse et une validation par un utilisateur humain. L'application de ces contraintes sur l'ensemble des motifs séquentiels fréquents est l'étape la plus coûteuse de notre méthode. La sélection des 65 000 motifs a pris 636 minutes.

La division de l'ensemble des motifs séquentiels obtenus à l'étape précédente en plusieurs sous-ensembles et la fouille récursive de ces sous ensembles permet de réduire encore de façon sensible le nombre de motifs séquentiels candidats pour représenter des interactions. La fouille récursive de chacun de ces sous-ensemble permet d'exhiber au plus k motifs séquentiels pour représenter ce sous-ensemble. Dans cette expérience, nous fixons le paramètre k à 4. Le

nombre de sous-ensembles créés est de 515 (365 pour les noms, 150 pour les verbes). A l'issu de la fouille récursive sur chaque sous-ensemble, il reste 667 motifs séquentiels susceptibles de représenter des interactions. Ce nombre sensiblement plus petit que les précédents garantit la faisabilité d'une analyse par un utilisateur humain. La fouille récursive de ces sous-ensembles est très peu coûteuse. Elle a duré environ 2 minutes.

Les 667 motifs séquentiels restants ont été analysés par deux utilisateurs. Ils ont validé 232 motifs séquentiels en 90 minutes. Ce qui signifie que ces 232 motifs séquentiels modélisent bien des interactions entre gènes (les motifs refusés étant porteurs d'autres relations sémantiques comme la modalité, la simple occurrence, etc). Parmi ces motifs, certains représentent explicitement des interactions comme les motifs *AGENE@np bind@vvz to@to AGENE@np .@sent*, *AGENE@np deplete@vvn AGENE@np .@sent* et *activation@nn of@in AGENE@np by@in AGENE@np .@sent* qui décrivent des interactions bien connues (liaison, inhibition, activation). D'autres motifs modélisent des interactions entre gènes de façon plus générale, signifiant simplement qu'un gène joue un rôle dans l'activité d'un autre gène comme les motifs *AGENE@np involve@vvn in@in AGENE@np .@sent*, *AGENE@np play@vvz role@nn in@in the@dt AGENE@np .@sent* et *AGENE@np play@vvz role@nn in@in of@in AGENE@np .@sent*.

Les motifs séquentiels obtenus forment des patrons linguistiques prêts à être appliqués dans des textes biomédicaux pour détecter des interactions entre gènes. Rappelons que pour être appliqués, ces patrons ne s'appuient sur aucune analyse syntaxique de la phrase. Il suffit de chercher à instancier chaque élément du patron dans la phrase.

3.2 Application des patrons linguistiques pour la détection d'interactions

Pour tester la qualité de nos patrons linguistiques, nous considérons trois jeux de données connus dans la littérature : *GeneTag* du jeu de données *Genia* (Tanabe *et al.*, 2005), *BioCreative* issu de (Yeh *et al.*, 2005), et *AIMed* de (Bunescu & Mooney, 2005). Dans ces jeux de données, les noms de gènes ou de protéines sont étiquetés. Dans chaque corpus, nous avons pris aléatoirement 200 phrases et testé si les patrons linguistiques s'appliquaient. Pour chaque phrase contenant une interaction, nous mesurons les performances des patrons linguistiques pour détecter ces interactions. Notons que nous avons également procédé à un étiquetage morpho-syntaxique de ces phrases pour pouvoir appliquer correctement les patrons linguistiques, de plus l'application des patrons linguistiques est quasi-instantanée.

Corpus	Précision	Rappel	F-Score
BioCreative (Yeh *et al.*, 2005)	$0,92$	$0,767$	$0,836$
GeneTag (Tanabe *et al.*, 2005)	$0,909$	$0,8$	$0,851$
AIMed (Bunescu & Mooney, 2005)	$0,93$	$0,84$	$0,88$

TAB. 1 – Tests menés sur différents corpus

Le tableau 1 décrit la précision, le rappel et le f-score ($\frac{2\times P\times R}{P+R}$) de l'application de patrons linguistiques sur chaque corpus. Les scores sont similaires sur les trois corpus. De plus, ces résultats sont encourageants dans la mesure où la précision est très bonne et le rappel satisfaisant. Ces résultats sont tout à fait comparables à ceux des autres méthodes présentes dans la littérature en notant toutefois que les tâches ne sont jamais identiques (Krallinger *et al.*, 2008).

3.3 Discussion

Bien que les outils d'analyse morpho-syntaxique présentent des résultats satisfaisants, il subsiste encore un nombre non négligeable d'erreurs d'étiquetage concernant la lemmatisation ou l'attribution d'une catégorie grammaticale. Notre méthode est assez robuste face à ce phénomène puisque ces erreurs sont également présentes lors de la découverte des motifs d'interactions. Ainsi, si une erreur est suffisamment fréquentes, elle sera présente dans un motif extrait. Par exemple, Tree Tagger ne lemmatise pas le mot *cotransfected* mais des motifs extraits contiennent la forme *cotransfected@vvn*.

Notons que la portée des patrons linguistiques se limitent au cadre de la phrase. Une telle portée peut introduire des ambiguïtés dans la détection d'interaction lorsque plus de deux gènes apparaissent dans la phrases. Plusieurs cas sont possibles. Soit plusieurs interactions binaires sont présentes dans la phrase, soit l'interaction est de type n-aire ($n \geq 3$) ou encore on peut trouver une interaction en présence d'une simple énumération de gènes. Le cas des interactions n-aires peut être résolu avec un apprentissage sur un jeux de données contenant des interactions n-aires. Les deux autres cas peuvent être traités en introduisant des règles limitant la portée des patrons, par exemple à l'aide de connecteurs (but, however, etc.).

4 Conclusion

Dans cet article, l'utilisation de l'extraction de motifs séquentiels permet de découvrir des patrons linguistiques pour détecter des relations entre entités nommées. Cette méthode est appliquée à la découverte d'interactions entre gènes. La découverte des patrons linguistiques est entièrement automatique et ne nécessite qu'une validation peu coûteuse en temps par un expert. En effet, le nombre de patrons linguistiques est limité par l'introduction de contraintes et un processus de fouille récursive. L'acquisition de ces patrons et leur application ne nécessitent ni analyse syntaxique ni ressource autre que le corpus d'apprentissage.

Nous pensons améliorer ce travail selon deux axes. D'une part, l'introduction de contraintes linguistiques lors de la phase de fouille permettra d'éviter l'application coûteuse de ces contraintes en post-traitement. La prise en compte de hiérarchie sur les items devrait permettre d'introduire de nouveaux types de contraintes et d'améliorer les motifs découverts, notamment par généralisation. D'autre part, nous envisageons la prise en compte des « modalités » qui interviennent souvent dans l'expression de l'interaction (négation, résultat d'expérience, impossibilité, etc.) et qui, de plus intéressent les biologistes. Parmi les motifs découverts non validés, bon nombre sont porteurs de modalités.

A terme, ce travail vise à être intégré dans une application plus large destinée aux biologistes. L'idée est de combiner plusieurs sources de connaissances (connaissances issues du transcriptome (Leyritz *et al.*, 2008) et des données textuelles issues de PubMed). En effet, croiser ces différentes sources devrait permettre de mettre en relief de nouvelles connaissances et d'introduire de nouvelles contraintes.

Remerciements : Nous remercions Christophe Rigotti pour nous avoir permis d'utiliser l'extracteur dmt4. Ce travail est partiellement financé par l'ANR, projet Bingo2[4] (ANR-07-MDCO-014).

[4] `http://bingo2.greyc.fr/`

Références

BUNESCU R. C. & MOONEY R. J. (2005). A shortest path dependency kernel for relation extraction. In *HLT/EMNLP* : The Association for Computational Linguistics.

KIM J.-H., MITCHELL A., ATTWOOD T. K. & HILARIO M. (2007). Learning to extract relations for protein annotation. In *ISMB/ECCB (Supplement of Bioinformatics)*, p. 256–263.

KRALLINGER M., LEITNER F., RODRIGUEZ-PENAGOS C. & VALENCIA A. (2008). Overview of the protein-protein interaction annotation extraction task of BioCreative II. *Genome Biology*, **9(Suppl 2)**.

LEYRITZ J., SCHICKLIN S., BLACHON S., KEIME C., ROBARDET C., BOULICAUT J.-F., BESSON J., PENSA R. & GANDRILLON O. (2008). SQUAT : A web tool to mine human, murine and avian SAGE data. *BMC Bioinformatics*, **9**(1), 378.

NANNI M. & RIGOTTI C. (2007). Extracting trees of quantitative serial episodes. In *Knowledge Discovery in Inductive Databases 5th Int. Workshop KDID'06, Revised Selected and Invited Papers*, p. 170–188 : Springer-Verlag LNCS 4747.

NG R. T., LAKSHMANAN L. V. S., HAN J. & PANG A. (1998). Exploratory mining and pruning optimizations of constrained association rules. In L. M. HAAS & A. TIWARY, Eds., *SIGMOD Conference*, p. 13–24 : ACM Press.

PEI J., HAN B., MORTAZAVI-ASL B. & PINTO H. (2001). Prefixspan : Mining sequential patterns efficiently by prefix-projected pattern growth. In *Proc. of the 17th Int. Conf. on Data Engineering (ICDE'01)*, p. 215–224.

ROSARIO B. & HEARST M. A. (2005). Multi-way relation classification : application to protein-protein interactions. In *HLT '05 : Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, p. 732–739, Morristown, NJ, USA : Association for Computational Linguistics.

SCHMID H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*.

SOULET A. (2007). Résumer les contrastes par l'extraction récursive de motifs. In *Actes de CAP'07, Conférence francophone sur l'apprentissage automatique - 2007, Grenoble, France*.

SRIKANT R. & AGRAWAL R. (1996). Mining sequential patterns : Generalizations and performance improvements. In P. M. G. APERS, M. BOUZEGHOUB & G. GARDARIN, Eds., *EDBT*, volume 1057 of *Lecture Notes in Computer Science*, p. 3–17 : Springer.

TANABE L., XIE N., THOM L., MATTEN W. & WILBUR J. (2005). GENETAG : a tagged corpus for gene/protein named entity recognition. *BMC Bioinformatics*, **6**, 10.

VETAH M. O. A., AUBIN S., BESSIÈRES P., BISSON G., HAMON T., LAGARRIGUE S., NAZARENKO A., NÉDELLEC C., POIBEAU T. & WEISSENBACHER D. (2004). Extraction d'information appliquée au domaine biomédical - apprentissage et taln. In *Actes de la Conférence International de Fouille de texte, CIFT-04. Marie Hélène Antoni et François Yvon (Eds.). La Rochelle.*

YEH A., MORGAN A., COLOSIMO M. & HIRSCHMAN L. (2005). BioCreAtIvE Task 1A : gene mention finding evaluation. *BMC Bioinformatics*, **6**, 10.

ZAKI M. J. (2001). Spade : An efficient algorithm for mining frequent sequences. *Machine Learning*, **42**(1/2), 31–60.

ZWEIGENBAUM P., DEMNER-FUSHMAN D., YU H. & COHEN K. B. (2007). Frontiers of biomedical text mining : current progress. *Brief Bioinform*, p. 358–375.