Fouille de motifs séquentiels pour la découverte de relations entre gènes et maladies rares

Nicolas Béchet[1], Peggy Cellier[2], Thierry Charnois[1], Bruno Cremilleux[1]

[1] GREYC, Université de Caen Basse-Normandie
Campus II science 3
14032 Caen CEDEX, France.
{nicolas.bechet, thierry.charnois, bruno.cremilleux}@unicaen.fr

[2] IRISA, INSA de Rennes/,
Campus de Beaulieu
35042 Rennes cedex, France
peggy.cellier@irisa.fr

Résumé : Orphanet est un organisme dont l'objectif est notamment de rassembler des collections d'articles traitant de maladies rares. Cependant, l'acquisition de nouvelles connaissances dans ce domaine est actuellement réalisée manuellement. Dès lors, obtenir de nouvelles informations relatives aux maladies rares est un processus chronophage. Permettre d'obtenir ces informations de manière automatique est donc un enjeu important. Dans ce contexte, nous proposons d'aborder la question de l'extraction de relations entre gènes et maladies rares en utilisant des approches de fouille de données, plus particulièrement de fouille de motifs séquentiels sous contraintes. Nos expérimentations montrent l'intérêt de notre approche pour l'extraction de relations entre gènes et maladies rares à partir de résumés d'articles de PubMed.

Mots-clés : Fouille de données, motifs séquentiels, extraction d'information, patrons linguistiques, maladies rares

1 Introduction

Les maladies rares sont l'un des enjeux prépondérant dans le domaine de la santé publique. Une maladie est considérée comme rare (notée MR, par la suite), si elle affecte moins de 1 personne sur 2000. Il existe actuellement entre 6 000 et 8 000 MRs référencées en Europe, affectant environ

30 millions de personnes et bien plus dans le reste du monde. Les informations relatives aux maladies rares sont éparpillées alors que les scientifiques ont besoin de les rassembler et de les partager afin de travailler plus efficacement. C'est pourquoi, Orphanet [1] propose une base de données internationale, accessible sur le Web, visant à rassembler une collection d'articles synthétiques rédigés par des experts sur les MRs. Cependant, la veille nécessaire sur la parution de nouveaux articles dans la littérature, et la relecture de ceux-ci sont des tâches actuellement réalisées manuellement. Celles-ci reposent sur des annotateurs humains, filtrant les articles traitant de maladies rares avec une cause génétique. Ainsi, produire une nouvelle documentation relative à une maladie rare est un processus fastidieux. L'acquisition automatique de connaissances liées aux maladies rares à partir d'une large collection de données textuelles est donc un enjeu particulièrement important. Dans ce contexte, nous nous sommes intéressés plus particulièrement au problème de l'extraction de relations de type gène–MR à partir de collections textuelles comme celle de PubMed (qui contient plus de 21 millions de publications biomédicales). Dans cet article, nous proposons d'aborder la question de l'extraction de relations entre gènes et maladies rares en utilisant des approches de fouille de données, plus particulièrement la fouille de motifs séquentiels sous contraintes.

La traitement automatique des langues (TAL), et l'extraction d'information en particulier, ont pour but de produire une analyse précise afin d'extraire des connaissances spécifiques telles que la reconnaissance d'entités nommées (par exemple un gène ou une maladie) et les relations entre les entités reconnues (par exemple les interactions entre gènes (Krallinger *et al.*, 2008) ou encore les relations entre maladies et traitements (Abacha & Zweigenbaum, 2011)). Ces approches issues du TAL nécessitent un certain nombre de règles comme des expressions régulières pour la recherche en surface (Giuliano *et al.*, 2006) ou des patrons syntaxiques (Rinaldi *et al.*, 2006; Fundel *et al.*, 2007). Bien qu'efficace, la définition manuelle de ces règles est une tâche chronophage et nous verrons comment les méthodes de fouille permettent de suggérer de telles règles.

Par ailleurs, les méthodes d'apprentissage telles que les SVM (Support Vector Machines) et les CRF (Conditional Random Fields) (Krallinger *et al.*, 2008), sont des processus automatiques et peu coûteux en termes de temps (et moins que les approches issues du TAL). Cependant, bien qu'obtenant généralement de bons résultats, ces approches ont leurs limites. En effet, les résultats produits par ces méthodes ne sont pas intelligibles par

1. www.orphanet.org

un humain. Celui-ci ne peut pas exploiter directement les résultats, par exemple pour former des patrons linguistiques pour des systèmes de TAL. En outre, ces méthodes d'apprentissage doivent disposer de corpus annotés dont l'acquisition est souvent coûteuse. De plus, ces corpus sont la plupart du temps spécifiques à un domaine (Hobbs & Riloff, 2010) et l'extraction d'information pour un nouveau problème nécessite alors l'acquisition d'un nouveau corpus.

Des travaux récents proposent de tirer bénéfice des avantages de chacune de ces approches et de combiner des techniques de TAL avec des méthodes de fouille de données. La fouille de données permet la découverte d'informations implicites à partir d'une collection de données (Frawley *et al.*, 1991). Ainsi, dans (Cellier *et al.*, 2010), une méthode est proposée pour extraire automatiquement des patrons linguistiques permettant la découverte de relations entres des entités nommées dans des corpus. Cette approche est non supervisée, et ne nécessite ni analyse syntaxique, ni ressource externe à l'exception d'un corpus d'apprentissage. Elle se fonde sur l'extraction de motifs séquentiels fréquents, où une séquence est une liste de littéraux appelés *items*, un item étant un mot (ou son lemme) provenant de nos données textuelles.

Cet article se situe dans la lignée de cette nouvelle voie prometteuse de l'hybridation fouille/TAL. Nous montrons l'intérêt des motifs séquentiels pour des tâches d'extraction d'information, et plus particulièrement dans le cadre de la découverte de relations entre gènes et maladies rares. Les motifs séquentiels extraits ne pouvant pas être directement exploités en raison de leur grand nombre, nous proposons de réduire leur nombre en intégrant des contraintes et en utilisant une représentation condensée des motifs séquentiels produits. De plus, l'approche que nous proposons est capable d'extraire et d'exploiter des motifs séquentiels composés d'*itemsets* et non plus de simples *items* comme dans (Cellier *et al.*, 2010). Concrètement, cela signifie qu'un mot peut être représenté par des informations multiples comme le mot lui même, son lemme, sa catégorie grammaticale. Il s'agit d'une contribution importante car de nombreuses applications nécessitent d'exprimer une information qui combine différents niveaux d'abstraction. Par exemple, nous présentons en section 3.2 l'intérêt d'un patron comme $\langle (mutation\, NNS)(IN)(isocitrate\, NN)\ (GENE)(be\, VBP)(DISEASE) \rangle$ [2] qui combine les niveaux d'abstraction lemme (en minuscules) et catégorie grammaticale (en majuscules). D'autre part, nous montrons l'apport des

2. Ce patron comporte un nom au pluriel (NNS) suivi d'une préposition (IN), du nom $isocitrate$, d'un nom de gène, du verbe be au présent et d'un nom de maladie

contraintes portant sur les motifs séquentiels afin d'adapter la recherche des motifs en fonction des intérêts de l'utilisateur, les contraintes permettant de filtrer des patrons linguistiques pertinents pour la découverte de relations entre gènes et maladies rares. Enfin, nous contribuons au domaine de la recherche d'information à partir de données textuelles, et plus précisément dans le cadre de la découverte de relations entre gènes et maladies rares.

Cet article est organisé comme suit. Nous introduisons dans un premier temps la notion de fouille de motifs séquentiels (section 2). Nous présentons ensuite notre méthode d'extraction de relations entre gènes et maladies rares à partir de textes biomédicaux (section 3). Finalement, les expérimentations décrites en section 4 montrent l'intérêt de la méthode pour la documentation relative aux maladies rares.

2 La fouille de motifs séquentiels

La fouille de motifs séquentiels (Agrawal & Srikant, 1995) est un champ de la fouille de données ayant pour but la découverte de régularités dans des données se présentant sous forme de séquences. Plusieurs algorithmes (Srikant & Agrawal, 1996; Yan *et al.*, 2003; Zaki, 2001) ont été proposés pour extraire les motifs séquentiels. Nous introduisons maintenant ce domaine qui est le support méthodologique des méthodes de découverte d'information que nous proposons dans cet article.

En fouille de données séquentielles, un *itemset*, noté $I = (i_1 \dots i_n)$ est un ensemble de littéraux appelés *items*. Par exemple, $(a\ b)$ est un itemset avec deux items a et b. Une séquence S est une liste ordonnée d'itemsets, notée $s = \langle I_1 \dots I_m \rangle$. Par exemple, $\langle (a)\ (a\ b\ c)\ (a\ c)\ (d) \rangle$ est une séquence de quatre itemsets. Une séquence $S_1 = \langle I_1 \dots I_n \rangle$ est dite *incluse* dans une autre séquence $S_2 = \langle I'_1 \dots I'_m \rangle$ s'il existe des entiers $1 \leq j_1 < \dots < j_n \leq m$ tels que $I_1 \subseteq I'_{j_1}$,..., $I_n \subseteq I'_{j_n}$. La séquence S_1 est alors appelée une sous-séquence de S_2, noté $S_1 \preceq S_2$. Par exemple, $\langle (a)(a\ c) \rangle$ est incluse dans $\langle (a)(a\ b\ c)(a\ c)(d) \rangle$.

Une base de séquences, notée SDB, est un ensemble de tuples (sid, S), où sid est un identifiant de séquence, et S est une séquence. Par exemple, la table 1 décrit une base de séquences composée de quatre séquences.

Un tuple (sid, S) *contient* une séquence S_1 si $S_1 \preceq S$. Le *support*[3]

3. Notons que le support relatif est également utilisé : $sup(S_1) = \frac{|\{(sid, S) \mid (sid, S) \in SDB \wedge (S_1 \preceq S)\}|}{|SDB|}$.

Identifiant	Sequence
1	$\langle(a)\ (a\ b\ c)\ (a\ c)\ (d)\rangle$
2	$\langle(a\ d)\ (c)\ (b)\rangle$
3	$\langle(a\ b)\ (d)\ (b)\ (c)\rangle$
4	$\langle(a\ d)\ (b)\ (b)\rangle$

TABLE 1 – Exemple de SDB.

d'une séquence S_1 dans une base de données de séquences SDB, noté $sup(S_1)$, est le nombre de tuples de la SDB contenant S_1. Par exemple, dans la table 1, $sup(\langle(a\ b)(c)\rangle) = 2$, car les séquences 1 et 3 contiennent $\langle(a\ b)(c)\rangle$. Un motif séquentiel ***fréquent*** est un motif ayant une support supérieur ou égal à un certain seuil $minsup$.

Dans la pratique, le nombre de motifs séquentiels fréquents peut être important. Une façon de réduire leur nombre sans perte d'information est l'utilisation des représentations condensées de motifs. En effet, à partir d'un ensemble de motifs spécifiques, comme les motifs séquentiels fermés (Yan ***et al.***, 2003), il est possible de régénérer, si on le souhaite, l'ensemble de tous les motifs séquentiels. De plus, les motifs fermés éliminent la redondance entre motifs supportés par un même ensemble de séquences, c'est pourquoi nous les utiliserons dans la suite de ce travail. Plus formellement, un motif séquentiel S est fermé s'il n'existe pas de motif séquentiel S' tel que $S \preceq S'$ et que $sup(S) = sup(S')$. Par exemple, avec $minsup = 2$, le motif séquentiel $\langle(a\ b)\rangle$ de la table 1 n'est pas fermé car $sup(\langle(a\ b)\rangle) = sup(\langle(a\ b)(c)\rangle)$ et $\langle(a\ b)\rangle \preceq \langle(a\ b)(c)\rangle$.

Une autre démarche complémentaire pour la découverte de motifs utiles, est l'utilisation de contraintes (Dong & Pei, 2007). Une contrainte permet de focaliser la recherche en fonction des centres d'intérêts de l'utilisateur et limite aussi le nombre de motifs séquentiels extraits. Un exemple très classique de contrainte, que nous avons déjà introduite, est celle de support minimum. Le paradigme de l'extraction de motifs séquentiels sous contraintes offre de multiples possibilités telle que par exemple la contrainte de gap. Un motif séquentiel avec contrainte de gap $[M, N]$, noté $P_{[M,N]}$ est un motif tel qu'au minimum M itemsets et au maximum N itemsets sont présents entre chaque itemset voisin du motif dans les séquences correspondantes. Par exemple, dans la table 1, $P_{[0,2]} = \langle(a)(c)\rangle$ et $P_{[1,2]} = \langle(a)(c)\rangle$ sont deux motifs séquentiels avec des contraintes de gap. $P_{[0,2]}$ est contenu dans trois séquences, les séquences 1, 2 et 3, tandis que $P_{[1,2]}$ est contenu dans seulement deux séquences, les séquences 1 et 3. En effet, dans la séquence 2, il n'y a pas d'itemset entre l'itemset contenant a et

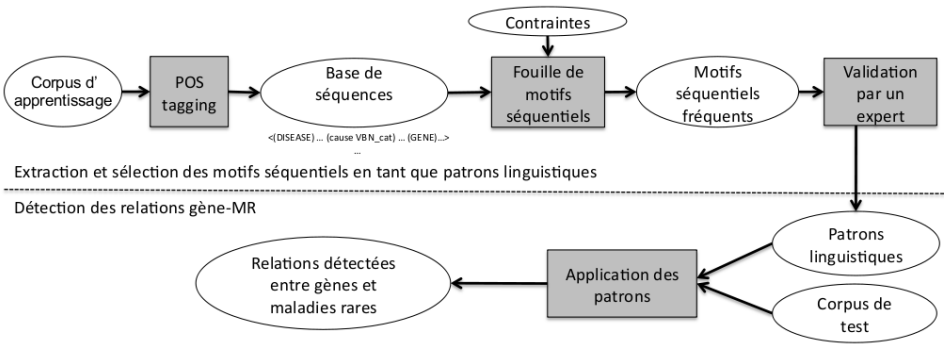

FIGURE 1 – Vue générale de la méthode d'extraction de relations gène–MR

l'itemset contenant c.

3 La découverte de relations entre gènes et maladies rares

Dans cette section, nous présentons le processus permettant la découverte de relations entres gènes et maladies rares (section 3.1). Puis, nous définissons les contraintes utilisées pour l'extraction des motifs séquentiels (section 3.2).

3.1 Vue générale de l'approche

La figure 1 présente le schéma global de l'approche. Celle-ci se décompose en deux parties : l'extraction et la validation de motifs séquentiels en tant que patrons linguistiques, et l'application de ces patrons afin de découvrir des relations entre gènes et maladies rares.

La base de séquences est construite à partir d'un corpus d'apprentissage. Les séquences sont les phrases du corpus d'apprentissage contenant au moins une maladie rare et un gène. Grâce à l'étape de POS tagging, chaque mot est remplacé par un itemset contenant le lemme du mot, et sa catégorie grammaticale [4]. La table 2 donne un extrait d'une base de sé-

4. Exemple de catégories grammaticales : DT : Déterminant, IN : Préposition or conjonction de subordination, JJ : Adjectif, NN : Nom, RB : Adverbe, VB : Verbe. La liste complète des catégories grammaticales est disponible ici http ://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

quences avec trois séquences [5]. Par exemple, dans S_1, le verbe "conclude" est remplacé par l'itemset $(conclude\ VBP)$, $conclude$ étant son lemme et VBP sa catégorie grammaticale (i.e. un verbe au présent et à une personne autre que la troisième). Les noms de gènes identifiés dans les phrases sont remplacés par un unique item *GENE* ; il en est de même pour les noms de maladies rares, remplacés par l'item *DISEASE*. Remarquons que contrairement aux méthodes d'apprentissage numérique, le corpus d'apprentissage utilisé pour notre approche ne contient pas de relations annotées.

Une fois la base de séquences constituée, les motifs séquentiels fréquents sont extraits de celle-ci (voir section 2), cette extraction se faisant sous des contraintes qui sont détaillées à la section suivante. Afin d'éviter la redondance entre motifs, nous utilisons les motifs séquentiels fermés fréquents (cf. section 2) et non pas les motifs séquentiels fréquents comme c'est le cas dans les approches existantes. D'autre part, un mot est représenté par un ensemble de descripteurs véhiculant différents types d'informations et non pas une information unique. Ce mode de représentation du mot permet de combiner différents niveaux d'abstraction et de produire aussi bien un motif séquentiel générique (c'est-à-dire ne contenant que des catégories grammaticales, comme $\langle(NNS)(IN)(NN)(GENE)\ (VBP)\ (DISEASE)\rangle$) ; qu'un motif plus spécifique comme $\langle(mutation\ NNS)(IN)(isocitrate\ NN)(GENE)(occur\ VBP)(DISEASE)\rangle$, motif combinant catégories grammaticales et lemmes.

Un expert effectue ensuite une sélection parmi les motifs séquentiels en ne conservant que ceux qu'il évalue comme des patrons linguistiques pertinents dans le cadre de la découverte de relations entre gènes et maladies rares. Pour faciliter l'exploration des motifs par l'expert, ceux-ci sont regroupés par verbe ou par nom (par exemple, tous les motifs contenant le verbe *occur* sont présentés en un ensemble).

Enfin, les patrons sélectionnés sont appliqués sur un nouveau corpus (corpus de test) afin d'en extraire de nouvelles relations gène–MR.

Identifiant de séquence	Séquence
S_1	$\langle(we\ PP)\ (conclude\ VBP)\ (that\ IN)\ (GENE)\ (be\ VBZ)$ $(essential\ JJ)\ (for\ IN)\ (maturation\ NN)\ (of\ IN)$ $(ubiquitin\ NN)\ (contain\ VBG)\ (autophagosomes\ NNS)$ $(and\ CC)\ (that\ DT)\ (defect\ NN)\ (in\ IN)\ (this\ DT)$ $(function\ NN)\ (may\ MD)\ (contribute\ VB)\ (to\ TO)$ $(DISEASE)\ (pathogenesis\ NN)\rangle$
S_2	$\langle(somatic\ JJ)\ (mutation\ NNS)\ (in\ IN)\ (isocitrate\ NN)$ $(dehydrogenase\ NN)\ (1\ CD)\ (GENE)\ (and\ CC)\ (GENE)$ $(occur\ VBP)\ (in\ IN)\ (glioma\ NNS)\ (and\ CC)\ (acute\ JJ)$ $(myeloid\ JJ)\ (leukaemia\ NN)\ (DISEASE)\rangle$
S_3	$\langle(DISEASE)\ (be\ VBZ)\ (normally\ RB)\ (cause\ VBN)\ (by\ IN)$ $(an\ DT)\ (autosomal\ JJ)\ (dominant\ JJ)\ (mutation\ NN)$ $(in\ IN)\ (the\ DT)\ (type\ NN)\ (i\ NN)\ (collagen\ NN)$ $(gene\ NNS)\ (GENE)\ (and\ CC)\ (GENE)\rangle$

TABLE 2 – Extrait de la base de séquences provenant de textes médicaux.

3.2 Contraintes pour modéliser des connaissances linguistiques

L'étape de fouille de motifs séquentiels s'effectue sous contraintes, cette stratégie permettant à la fois de modéliser des connaissances linguistiques et de filtrer les motifs les plus pertinents en fonction de la problématique. Le but est d'obtenir des motifs séquentiels qui traduisent certaines régularités linguistiques (c.-à-d. des relations gène–MR). Nous avons déjà introduit certaines contraintes classiques de la littérature, comme les contraintes de support minimum, de gap ou encore "être un motif fermé". Nous définissons maintenant d'autres contraintes qui s'avéreront précieuses pour la découverte de relations entre des gènes et des maladies rares dans les textes.

La contrainte d'*appartenance* permet de ne conserver que les motifs séquentiels contenant certains items sélectionnés. Par exemple, nous pouvons imposer qu'un motif séquentiel contienne les items : $GENE$ et $DISEASE$. La contrainte de *longueur minimum* ($minlength$) conserve uniquement les motifs séquentiels qui ont une taille supérieure à un seuil défini en nombre d'itemsets, il est alors simple de ne garder, par exemple,

5. S_1. *"We conclude that VCP is essential for maturation of ubiquitin-containing autophagosomes and that defect in this function may contribute to IBMPFD pathogenesis."*
S_2. *"Somatic mutations in isocitrate dehydrogenase 1 (IDH1) and IDH2 occur in gliomas and acute myeloid leukaemia (AML)."*
S_3. *"Osteogenesis imperfecta is normally caused by an autosomal dominant mutation in the type I collagen genes COL1A1 and COL1A2."*

que les motifs contenant au moins 3 mots. La contrainte de portée ($scope$) permet de définir le nombre maximum d'itemsets existant entre le premier itemset et le dernier itemset d'un motif dans les séquences de la base contenant le motif.

La contrainte d'*association* associe un type d'item à un autre type d'item. Par exemple, elle permet d'exprimer que tous les motifs séquentiels contenant l'item VB représentant la catégorie grammaticale "verbe" doivent également contenir un lemme (le lemme du verbe en question) dans le même itemset. Ainsi, le motif $\langle(GENE)(VB)(DISEASE)\rangle$ ne respecte pas cette contrainte d'association, alors que le motif $\langle(GENE)(coder\ VB)(DISEASE)\rangle$ est correct.

Il existe dans la littérature de nombreux algorithmes pour extraire des motifs séquentiels fréquents (par exemple (Srikant & Agrawal, 1996; Zaki, 2001)) ou encore des motifs fermés (par exemple (Yan *et al.*, 2003)). Cependant, à notre connaissance, il n'existe pas d'algorithme permettant d'extraire des motifs séquentiels fermés d'itemsets intégrant les différentes contraintes que nous venons d'énoncer. C'est pourquoi nous avons conçu un nouvel algorithme d'extraction de motifs séquentiels capable de prendre en compte, au sein même du processus de fouille, les contraintes en fonction de leurs propriétés (*e.g.* monotonie, anti-monotonie). Cet algorithme ne faisant pas l'objet de cet article, nous ne le décrivons pas ici.

4 Expérimentations

4.1 Protocole expérimental

Nous avons construit un corpus à partir de la base de données PubMed. Pour cela, nous avons interrogé cette base de telle sorte qu'elle retourne des articles relatifs à au moins un nom de gène provenant du dictionnaire HUGO et une maladie rare provenant du dictionnaire d'Orphanet. Nous avons conservé parmi ces articles les phrases contenant un gène et une maladie rare identifiés dans les dictionnaires HUGO et Orphanet. Les noms des maladies rares et les noms de gènes ont ensuite été généralisés. Par exemple, la phrase "*<**disease**>**Muir-Torre syndrome**<**disease**> is usually inherited in an autosomal dominant fashion and associated with mutations in the mismatch repair genes, predominantly in <**gene**>**MLH1**<**gene**> and <**gene**>**MSH2**<**gene**> genes.*" contient un nom de maladie rare identifié, et deux noms de gènes identifiés. Finalement, 17 527 phrases sont obtenues. À partir de celles-ci, nous avons tiré aléatoirement 200 phrases, qui définissent un corpus de

test, les autres phrases formant le corpus d'apprentissage.

Les 200 phrases du corpus de test ont été évaluées par un expert afin de définir les phrases contenant des relations entre gènes et maladies rares (189 relations ont été identifiées). Notons qu'il est possible qu'une phrase de ce corpus contienne plusieurs relations gène–MR, ou bien aucune.

4.1.1 L'extraction des motifs séquentiels.

Les séquences de la SDB sont les phrases du corpus d'apprentissage. Ces dernières sont construites comme décrit à la section 3.2. Les catégories grammaticales sont obtenues en utilisant l'outil TreeTagger (Schmid, 1994).

Nous avons expérimenté différents paramètres portant sur les contraintes :

- *le support minimal*. Trois valeurs de seuil sont utilisés : 0,5% (88 séquences), 0,2% (35 séquences), et 0,05% (8 séquences).
- *la contrainte de gap* : extraction avec et sans contrainte de gap (cette dernière étant fixée empiriquement à [0,10]).
- *la portée maximale*. Nous avons fixé la portée maximale à 20 afin de réduire le nombre de patrons extraits. Ainsi, le nombre maximal d'itemsets existant entre le premier itemset et le dernier itemset d'un motif possédant une relation de type gène–MR ne peut excéder 20 (c.-à-d. 20 mots dans les séquences de la SDB qui contiennent le motif).
- *la longueur minimale* : extraction avec et sans cette contrainte. La valeur de la longueur minimale a été fixée à 4. Cette valeur est un bon compromis pour limiter le nombre de patrons génériques obtenus.
- *la contrainte d'appartenance*. Les motifs séquentiels extraits doivent tous contenir au moins trois items : un gène, une MR, et un nom ou un verbe (en effet, ce type de mots est nécessaire pour caractérisant les relations recherchées).
- *la contrainte d'association*. Nous imposons que chaque nom et chaque verbe composant un motif soient associés à un lemme dans le même itemset (et non pas uniquement la catégorie grammaticale).

Par ailleurs, nous ne nous intéressons dans notre cas d'étude qu'aux relations binaires contenant au maximum un gène et au maximum une MR.

4.1.2 L'application des patrons linguistiques.

Une fois extraits et validés, les motifs sont appliqués en tant que patrons linguistiques sur le corpus de test. Ces patrons ne s'appuient sur aucune

minsup	gap	longueur min	nb. motifs	nb. motifs validés
0,50%	[0,10]	all	24 888	6 346
0,50%	[0,10]	4	22 794	6 310
0,50%	no gap	all	23 823	6 193
0,50%	no gap	4	22 084	6 156
0,20%	[0,10]	all	133 533	54 512
0,20%	[0,10]	4	126 777	54 429
0,20%	no gap	all	138 175	56 404
0,20%	no gap	4	130 579	56 290
0,05%	[0,10]	all	1 530 085	416 786
0,05%	[0,10]	4	1 493 914	416 533

TABLE 3 – Nombre de motifs séquentiels extraits en fonction des contraintes de gap et de support minimum.

analyse syntaxique : il suffit de chercher à instancier chaque élément du patron au sein de la phrase. Les contraintes de gap et de portée maximale sont également utilisées lors de l'application des patrons : pour le gap, les valeurs sont "pas de gap" ou une valeur de gap de [0,10] ; pour la portée maximale, "pas de portée maximale" ou une valeur de 20. Nous nommons ces contraintes "application gap" et "application scope".

4.2 Résultats

Nous présentons dans la table 3 le nombre de motifs séquentiels extraits en fonction des différentes valeurs des contraintes utilisées pendant l'extraction (colonne *Nb motifs*), le nombre de motifs séquentiels validés par l'expert (colonne *Nb motifs validés*). La contrainte de support minimal ($minsup$) est la plus influente en terme de nombre de motifs extraits [6].

Les résultats présentés en table 4 montrent l'impact de la contrainte de gap et du support minimal. Le meilleur f-score est obtenu avec le plus faible support minimum (0,05%) et la contrainte de gap à [0,10]. L'influence de la contrainte de gap avec la valeur [0,10] n'est pas la même en fonction des différents seuils de support minimum. Ainsi, avec $minsup$=0.50%, les résultats sont de meilleure qualité sans gap, alors

6. Notons que le nombre de motifs obtenus avec gap peut être supérieur au nombre de motifs obtenu sans gap. Ce phénomène, qui peut paraître surprenant, est dû au calcul de la fermeture des motifs qui est effectué après l'application de la contrainte de gap.

minsup	gap	rappel	précision	F-score
0,50%	[0,10]	0,37	0,67	0,48
0,50%	no gap	0,46	**0,69**	0,55
0,20%	[0,10]	0,50	0,65	0,56
0,20%	no gap	0,53	0,64	0,58
0,05%	[0,10]	**0,65**	0,66	**0,65**

TABLE 4 – Résultats expérimentaux obtenus pour différentes valeurs de support minimal et de gap.

minsup	longueur min.	rappel	précision	F-score
0,50%	all	0,37	0,67	0,48
0,50%	4	0,36	0,68	0,47
0,20%	all	0,50	0,65	0,56
0,20%	4	0,48	0,67	0,56
0,05%	all	0,65	0,66	0,65
0,05%	4	0,64	0,66	0,65

TABLE 5 – L'impact de la contrainte de longueur minimale.

qu'avec $minsup$=0,20%, la précision est légèrement améliorée. Nous expliquons ces différences par le nombre de motifs séquentiels extraits en fonction du support minimal (cf table 3). En effet, pour un seuil de support minimal élevé, très peu de motifs sont extraits (environ 6 000 pour $minsup$=0,50%). Sans contrainte de gap, nous obtenons plus de motifs génériques, ce qui peut substantiellement améliorer le rappel comme le montre les résultats de la table 4 avec $minsup$=0,50%. Une valeur élevée de $minsup$ améliore la précision alors qu'une plus petite valeur de $minsup$ améliore le rappel.

La table 5 montre l'impact de la contrainte de longueur minimale : celle-ci améliore un peu la précision. La table 6 donne les résultats sur l'étude des contraintes *application gap* et *application portée* lors de l'application des patrons. Dans cette table, le gap utilisé lors de l'extraction des patrons est [0,10], et la longueur minimale est fixée à 20. L'utilisation de cette dernière contrainte lors de l'application des patrons dégrade fortement le rappel, et également le f-score. En revanche, la contrainte de gap, toujours lors de l'application des patrons, améliore légèrement la précision, mais dégrade le rappel.

minsup	application gap	application portée	rec.	prec.	F-score
0,50%	[0,10]	all	0,33	0,68	0,44
0,50%	[0,10]	20	0,25	0,68	0,37
0,50%	no gap	all	0,37	0,67	0,48
0,50%	no gap	20	0,26	0,68	0,37
0,20%	[0,10]	all	0,48	0,66	0,55
0,20%	[0,10]	20	0,35	0,66	0,46
0,20%	no gap	all	0,50	0,65	0,56
0,20%	no gap	20	0,36	0,66	0,46
0,05%	[0,10]	all	0,60	0,66	0,63
0,05%	[0,10]	20	0,41	0,65	0,50
0,05%	no gap	all	0,65	0,66	0,65
0,05%	no gap	20	0,41	0,65	0,50

TABLE 6 – Résultats obtenus avec les contraintes *application gap* et *application portée* lors de l'application des patrons.

Contraintes	**rappel**	**précision**
freq. min.	Améliore	Pas d'effet
longueur min.	Dégrade	Améliore
gap	Dégrade	Pas d'effet
application gap	Dégrade	Pas d'effet
application portée	Dégrade	Pas d'effet

TABLE 7 – L'impact des contraintes utilisées sur la découverte de relations gène-MR.

La section suivante discute des résultats obtenus.

4.3 Discussion

La table 7 résume l'impact des différentes contraintes utilisées sur la qualité des résultats. Le support minimal et la longueur minimale sont les contraintes les plus pertinentes pour notre application, améliorant respectivement le rappel et la précision. En effet, si l'on souhaite favoriser la précision, il faut contraindre la taille des motifs extraits. Inversement, si l'on souhaite favoriser le rappel, il faut choisir une petite valeur pour le support minimum.

La précision maximale obtenue avec notre approche est de 0,69 et le

meilleur rappel est de 0,65. Ces résultats sont proches de ceux d'autres méthodes de la littérature pour des tâches similaires telles que (Abacha & Zweigenbaum, 2011) mais où les patrons sont construits manuellement. Rappelons que notre méthode découvre automatiquement les motifs et que l'utilisation des contraintes permet d'orienter la recherche sur les centres d'intérêts de l'utilisateur. La méthode a mis en évidence des patrons pertinents pour la découverte de relations gène–MR comme par exemple : $\langle(DISEASE)(be\ \ VBP))(JJ)(IN)(factor\ \ NN)(GENE)\rangle$, $\langle(DISEASE)(be\,VBZ))(JJ)(DT)(dominant\ JJ)(cause\,VBN)(by\,IN)$ $(DT)(GENE)\rangle$, ou encore $\langle(JJ)(DISEASE)(superoxide\,NN)(dismutase$ $NN)(GENE)\rangle$.

Nous effectuons maintenant une analyse qualitative des erreurs relevées. Certains faux négatifs (influençant la valeur de rappel) sont expliqués par le fait que les experts qui ont validé les motifs séquentiels comme patrons linguistiques, se sont focalisés sur la notion de causalité (c.-à-d. gène cause MR). Ainsi, une phrase comme "*We report on a case of B-ALL of L3 morphology with MYC- IGH translocation*" n'est pas découverte avec nos patrons, car elle contient un nombre important de termes génériques qui ne reflètent pas cette notion de causalité. Citons par exemple "report", "case" ou encore "morphology".

Certains cas de faux positifs, réduisant la précision, peuvent être expliqués par des erreurs lors de l'étape de reconnaissance d'entités nommées. Par exemple des noms de gènes ont été reconnus comme des noms de maladies rares. Citons la phrase "*One of the most versatile defence mechanisms against the accumulation of DNA damage is nucleotide excision repair, in which, among others, the Xeroderma pigmentosum group C (XPC) and group A (XPA) proteins are involved.*". Dans cet exemple, "Xeroderma pigmentosum" a été identifié comme une maladie alors qu'il s'agit d'un nom de gène. Il est possible que de faux positifs apparaissent à cause d'une erreur d'expertise dans le corpus de test. En effet, certaines phrases de ce corpus ont été jugées négatives par l'expert, alors qu'elles contenaient une relation gène–MR. Citons par exemple la phrase "*Small granular SOD1-immunoreactive inclusions were found in spinal motoneurons of all 37 sporadic and familial ALS patients studied, but only sparsely in 3 of 28 neurodegenerative and 2 of 19 non-neurological control patients.*" qui a été jugée négative.

La négation est un problème classique en TAL et certains faux positifs s'expliquent par celle-ci. Par exemple, la phrase "*None of these patients had ATP13A2 sequence variants likely to be causal for their disease, sug-*

gesting that mutations in this gene are not common causes of Kufs disease" est extraite lors de l'application de nos patrons, ce qui signifie qu'elle est jugée comme contenant une relation gène-MR, alors que l'expert l'a exclue.

5 Conclusion

Nous avons proposé dans cet article une nouvelle approche fondée sur l'extraction de motifs séquentiels afin de découvrir automatiquement des relations entre des gènes et des maladies rares à partir de textes biomédicaux. Les motifs séquentiels extraits sont utilisés comme des règles d'extraction d'information. Notre approche utilise un corpus où des gènes et des maladies rares sont identifiés mais ne nécessite pas de connaître au préalable les relations gène–MR de ce corpus. De plus, les patrons linguistiques produits sont intelligibles par un humain qui peut aisément les modifier ou les exclure si besoin.

Nous avons validé notre approche dans le cadre de la découverte de relation de type gène–MR, à partir de résumés d'articles de PubMed. Les résultats obtenus montrent l'intérêt de notre approche, le rôle des différentes contraintes utilisées pour extraire les motifs, et finalement conduisent à améliorer la documentation sur les maladies rares.

Remerciements

Les auteurs tiennent à remercier Marie-Christine Jaulent et l'équipe Orphanet INSERM pour la mise à disposition des données sur les maladies rares et les discussions très fructueuses.

Ce travail bénéficie du soutien de la région Basse-Normandie et de l'ANR (projet Hybride ANR-11-BS02-002).

Références

ABACHA A. B. & ZWEIGENBAUM P. (2011). A hybrid approach for the extraction of semantic relations from medline abstracts. In *Computational Linguistics and Intelligent Text Processing*, LNCS, p. 139–150 : Springer.

AGRAWAL R. & SRIKANT R. (1995). Mining sequential patterns. In *Int. Conf. on Data Engineering* : IEEE.

CELLIER P., CHARNOIS T. & PLANTEVIT M. (2010). Sequential patterns to discover and characterise biological relations. In *Computational Linguistics and Intelligent Text Processing*, LNCS, p. 537–548 : Springer.

DONG G. & PEI J. (2007). *Sequence Data Mining*. Springer.

FRAWLEY W. J., PIATETSKY-SHAPIRO G. & MATHEUS C. J. (1991). Knowledge discovery in databases : An overview. In *Knowledge Discovery in Databases*. AAAI/MIT Press.

FUNDEL K., KÜFFNER R. & ZIMMER R. (2007). RelEx - relation extraction using dependency parse trees. *Bioinformatics*, **23**(3), 365–371.

GIULIANO C., LAVELLI A. & ROMANO L. (2006). Exploiting shallow linguistic information for relation extraction from biomedical literature. In *Proc. of the Conf. of the European Chapter of the Association for Computational Linguistics* : The Association for Computer Linguistics.

HOBBS J. R. & RILOFF E. (2010). Information extraction. In N. INDURKHYA & F. J. DAMERAU, Eds., *Handbook of Natural Language Processing, Second Edition*. Boca Raton, FL : CRC Press, Taylor and Francis Group.

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

RINALDI F., SCHNEIDER G., KALJURAND K., HESS M. & ROMACKER M. (2006). An environment for relation mining over richly annotated corpora : the case of genia. *BMC Bioinformatics*, **7**(S-3).

SCHMID H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proc. of the Int. Conf. on New Methods in Language Processing*, Manchester, UK.

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 *LNCS*, p. 3–17 : Springer.

YAN X., HAN J. & AFSHAR R. (2003). Clospan : Mining closed sequential patterns in large databases. In D. BARBARÁ & C. KAMATH, Eds., *SDM* : SIAM.

ZAKI M. J. (2001). SPADE : An efficient algorithm for mining frequent sequences. *Machine Learning Journal*, **42**(1/2), 31–60. special issue on Unsupervised Learning.