Enrichissement automatique d'une base de connaissances biologiques à l'aide des outils du Web sémantique

Ines Jilani[1], Florence Amardeilh[2]

[1]INSERM, UMR S 872, Éq. 20, Les Cordeliers, Paris, F-75006 France ; Université Pierre et Marie Curie-Paris6, UMR S 872, Paris, F-75006 France ; Université Paris Descartes, UMR S 872, Paris, F-75006 France,
ines.jilani@spim.jussieu.fr
[2]Modyco, UMR 7114, Université Paris 10, F-92001 Nanterre Cedex, France
florence.amardeilh@mondeca.com

Résumé : Collecter, lire, interpréter et annoter une grande masse de données textuelles n'est pas chose facile depuis le développement des nouvelles technologies dont Internet qui propose pléthore d'informations. Ces tâches sont d'autant plus fastidieuses à mener dans le domaine de la biologie où les intervenants doivent constamment être informés des nouveautés mais aussi réaliser des expériences sur la paillasse pour publier à leur tour leurs travaux et rester concurrentiels. Cet article propose de construire une ontologie, de la peupler automatiquement grâce à une méthode de traitement automatique des langues : les patrons lexico-syntaxiques. Une évaluation de l'extraction de connaissances est réalisée et présente une précision de 72% ainsi qu'un rappel de 50%.

Mots-clés : Ingénierie des connaissances, Apprentissage machine, Jeunes chercheurs en Intelligence Artificielle, Peuplement ontologique.

1 Introduction

Les nouvelles techniques apportées par les progrès dans le domaine de la biologie moléculaire comme le séquençage des génomes, ou les puces à ADN produisent une masse de données si grande que les biologistes ont du mal à s'y retrouver. L'accès à ces données mais aussi leur interprétation ainsi que leurs annotations sont des tâches trop fastidieuses à réaliser avec les outils classiques du biologiste.

Une autre difficulté vient s'ajouter à l'augmentation du volume des données, il s'agit des nombreuses terminologies employées par les biologistes : chacune est établie pour un sous domaine précis de la biologie. Or, un même terme peut parfois avoir des significations différentes selon que l'on s'intéresse à un sous domaine ou à un autre. Par ailleurs, un grand nombre de bases de données spécialisées voient le jour et sont accessibles à tous sur la toile. Cependant elles utilisent généralement des entrées différentes, mais aussi des terminologies propres à chacune. Ainsi, les

biologistes ne peuvent avoir une vision globale d'une connaissance disponible à partir d'une seule et même source car il n'existe pas de base de données qui mutualise toutes celles disponibles en biologie.

Nous exposons tout d'abord un bref état de l'art concernant le domaine de travail et les méthodes utilisées, puis nous introduisons le contexte de l'étude. Ensuite, nous présentons la modélisation de l'ontologie, l'acquisition des connaissances sur les micro Acides RiboNucléiques (miARN) puis son peuplement effectif grâce aux connaissances extraites automatiquement avant de finir par conclure.

2 L'apport du Web Sémantique

Le Web Sémantique consiste à décrire le contenu de ses ressources en les annotant avec des informations non ambiguës afin de favoriser l'exploitation de ces ressources par des agents logiciels (Prié & Garlatti, 2004). Or, les données actuelles du Web sont souvent écrites en langage naturel, car destinées aux humains. Le langage naturel étant par essence trop ambigu, des alternatives formelles et sémantiquement explicites doivent être mises en place pour lever les ambiguïtés du langage naturel, aussi bien dans le contenu des ressources que dans leurs annotations.

Les ontologies, originaires des techniques de modélisation de la connaissance notamment développées en intelligence artificielle, sont utilisées dans le domaine de la biologie afin de proposer un ensemble structuré de tous les termes représentant le sens d'un champ d'information, une sorte de description de la structure des informations disponibles sur le sujet. L'ontologie constitue en soi un modèle représentatif de l'ensemble des concepts dans le sous domaine de la biologie concerné, ainsi que les relations entre ces concepts. Autrement dit, elle fournit les moyens d'exprimer les concepts d'un domaine en les organisant hiérarchiquement et en définissant leurs propriétés dans un langage de représentation des connaissances formel favorisant le partage d'une vue consensuelle sur ce domaine entre les applications informatiques qui en font usage (Bourigault, Aussenac-gilles, & Charlet, 2004).

L'exploitation des outils du Web Sémantique, et notamment l'exploitation des ontologies du domaine pour la tâche d'enrichissement automatique de bases de connaissances est encore innovante. Néanmoins il existe des systèmes (KEOPS (Brisson & Collard, 2007), (Hignette, 2007)) qui s'intéressent à l'extraction d'informations, à l'annotation sémantique, basées sur des ontologies de domaine ou non.

La mise en œuvre du peuplement de telles ontologies grâce aux solutions proposées par le Web sémantique passe par le traitement du langage naturel, complété par une ontologie de référence. Dans ce cas, la tâche consiste à repérer dans le texte les instances, existantes ou nouvelles, de cette ontologie. Lorsque de nouvelles instances sont identifiées dans le texte, extraites puis reliées à l'ontologie, il s'agit alors de peupler cette ontologie, c.-à-d. d'enrichir la base de connaissances y étant associée avec ces nouvelles instances.

Le peuplement de notre ontologie est réalisé grâce aux connaissances extraites automatiquement avec la méthode des patrons lexico-syntaxiques (PLS) (Jilani, Grabar, & Jaulent, 2006), issue du domaine du traitement automatique des langues et de la théorie des automates. La méthode des PLS a été utilisée pour la structuration automatique de terminologies : avec la détection de relations hiérarchiques et transversales. Cette méthodologie a été proposée la première fois par Hearst (Hearst, 1992) pour l'acquisition d'hyponymes à partir de textes. Elle a ensuite été développée par Morin dans le cadre de l'acquisition des patrons pour le repérage de relations hiérarchiques entre termes (Morin, 1999).

3 Contexte

Notre travail a pour point de départ un besoin spécifique des biologistes de l'Institut Pasteur (IP) de Montevideo (Uruguay). En effet, travaillant sur les miARN de l'espèce humaine, les biologistes ont besoin de récolter un maximum de connaissances liées à ce sujet. Les miARN sont des ARN simple-brin longs d'environ 21 à 24 nucléotides et sont des répresseurs post-transcriptionnels : en s'appariant à des ARN messagers, ils guident leur dégradation, ou la répression de leur traduction en protéine, entraînant l'apparition ou au contraire l'inhibition de maladies. Or, ce champ d'étude est très récent dans le domaine de la biologie, faisant même l'objet d'un prix Nobel en 2006. Par conséquent, beaucoup de chercheurs biologistes se sont penchés sur cette nouvelle problématique et ont mené un certain nombre de travaux ces dernières années afin de découvrir les interactions possibles entre des combinaisons de couples miARN et ARN messager. Du fait de son caractère novateur et du nombre de travaux récemment publiés, les biologistes possèdent une vision très restreinte sur le sujet. Nous pensons qu'un moyen de les aider à recenser les connaissances existantes, les différentes expérimentations menées, leurs résultats ainsi que les terminologies utilisées dans ce domaine, est de développer une application Web sémantique reposant sur une ontologie de ce domaine. La modélisation d'une telle ontologie permettra de fournir un moyen d'accès central aux différentes terminologies, aux instances de la base de connaissances voire même aux ressources des bases de données externes.

Mais cela ne suffit pas à la tâche des biologistes, il faut aussi leur fournir le moyen de mettre à jour cette base de connaissances sur les miARN. Pour cela, nous avons construit une plateforme basée sur un outil d'extraction d'informations, miR Discovery. L'alimentation de la base de connaissances par les annotations générées à partir des articles biologiques disponibles via le portail PubMed de Medline[1] sera exploitée afin d'appliquer dans un second temps des règles d'inférence qui permettront de raisonner sur ces miARN. Nous allons à présent décrire l'ontologie du domaine des miARN que nous avons construite manuellement avant d'aborder le problème de son alimentation automatique.

[1] www.ncbi.nlm.nih.gov/entrez/

4 Modélisation d'une ontologie des miARN

S'il n'existe pas à l'heure actuelle d'ontologie représentant la connaissance au sujet des miARN, plusieurs ressources terminologiques et ontologiques (Gene ontology (Ashburner et al., 2000), Sequence Ontology (Eilbeck et al., 2005), etc.) offrent un point de vue général sur les gènes et leur séquençage. De plus, des bases de données sur les ARNm et les miARN, telles que Tarbase (Sethupathy, Corda, & Hatzigeorgiou, 2006) et miRBase (Griffiths-Jones, 2004), ont vu le jour récemment, témoignant ainsi de l'engouement des biologistes pour cette nouvelle problématique. Pour ce projet, nous avions besoin d'une ontologie de haut niveau qui constituerait un bon point de départ pour élaborer une nouvelle ontologie dédiée à la représentation des miARN et de leurs impacts sur la régulation et la mutation des gènes. Nous avons donc choisi de travailler avec la Sequence Ontology (SO) car elle a pour objectif de décrire les séquences biologiques en général. Les concepts de gène, d'ARNm et de miARN entre autres étaient déjà représentés dans la SO ainsi que certaines relations pertinentes pour notre domaine comme « *is_part_of* » pour décrire la décomposition d'un miARN en « *loop* », « *stem* », « *3'UTR* », « *5'UTR* », etc. Nous avons également réutilisé une autre relation intéressante « *regulated_by* » pour modéliser le phénomène de régulation entre un miARN et un segment d'ARNm bien qu'elle possède à présent le statut « *deprecated* » dans la SO. Enfin, nous avons enrichi et étendu la modélisation existante de la SO avec la connaissance actuelle des miARN détenue par les biologistes de l'IP de Montevideo.

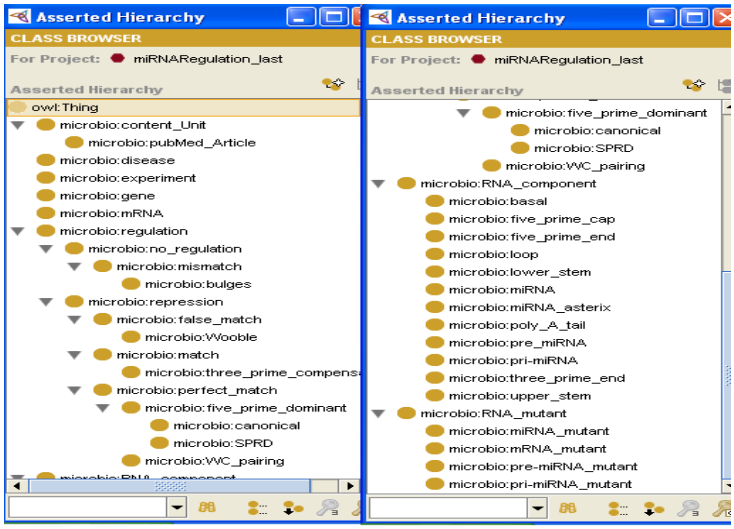

Fig. 1 – Taxonomie de l'ontologie des miARN dans Protégé

Plusieurs versions de la Sequence Ontology existent dans différents formats. Nous avons utilisé le format OBO et l'éditeur OBO Edit[2] pour visualiser cette ontologie et rechercher les concepts pertinents pour notre ontologie. Mais pour la

[2] http://oboedit.org/

modélisation de notre ontologie, nous utilisons le format OWL[3] DL via l'éditeur d'ontologie Protégé[4] largement utilisé par la communauté du Web Sémantique. Ce format OWL nous permet notamment de conserver un lien vers les concepts réutilisés de la SO et de faciliter l'intégration à venir de notre ontologie dans la SO. Lorsqu'un concept provient de la SO ou est équivalent à un concept existant de la SO, nous créons un lien sémantique entre le concept de notre ontologie et celui de la SO via la construction "*owl:equivalentClass*". Par exemple, le concept "miRNA" de la SO est une classe équivalente au concept "miARN" de notre ontologie. Nous appliquons le même principe pour les équivalences entre propriétés via la construction "*owl:equivalentProperty*". Le fait d'opter dans un premier temps pour une modélisation séparée de la SO nous permet de garder une indépendance de conception de notre ontologie tant qu'elle n'aura pas été définitivement validée par les biologistes, tout en conservant un lien fort avec la SO pour le jour où nous leur transmettrons nos résultats pour une demande d'intégration avec leurs ressources.

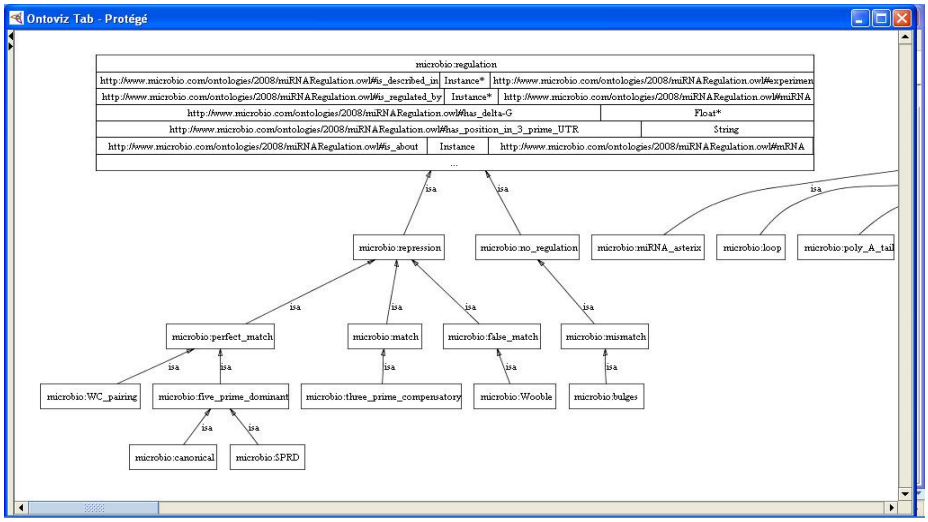

Fig 2 – Concepts, relations et attributs qui ont été ajoutés à la SO afin de représenter les différentes régulations induites par les miARN. La première version de l'ensemble de l'ontologie miARN est actuellement en cours de validation par les biologistes de l'IP.

La figure 1 présente la taxonomie de l'ontologie des miARN. L'ontologie représente aujourd'hui :

- 38 concepts primitifs : ils représentent les objets, abstraits ou concrets, réels ou fictifs, élémentaires ou composites, du monde réel. Ces concepts sont organisés en taxonomie, par l'utilisation de la relation de subsumption, dans laquelle ils peuvent appartenir à plusieurs sur-concepts différents. Par exemple, la classe "3_prime_UTR" est une sous-classe de "RNA_component".

[3] http://www.w3.org/2004/OWL/

[4] http://protege.stanford.edu/

- 13 relations (object properties) : elles représentent des interactions entre concepts permettant de construire des représentations complexes de la connaissance du domaine. Dans le domaine modélisé, les concepts « Disease » et « RNA_mutant » sont reliés entre eux par la relation sémantique « is_produced_by(Disease, RNA_mutant) » dans laquelle « Disease» est le domaine et « RNA_mutant » la portée (ou « range » en anglais).
- 11 attributs (datatype properties) : les **attributs** correspondent à des caractéristiques, des spécificités particulières attachées à un concept et qui permettent de le définir de manière unique dans le domaine. Leurs valeurs sont littérales, c.-à-d. de type primitif, comme une chaine de caractère ou un nombre entier.

Les **instances** de concepts ne font pas partie à proprement parler de l'ontologie, mais plutôt de la base de connaissances (Handschuh, 2005). En effet, ces dernières permettent de stocker les instances des concepts, mais aussi les instances de relations et les valeurs des propriétés en fonction des contraintes imposées par l'ontologie. C'est l'automatisation de cette tâche à partir d'un corpus d'articles de la littérature biologique que nous développons dans le reste de cet article.

5 Acquisition de connaissances sur les miARN

Afin de peupler cette ontologie du domaine et d'aider à la constitution des ressources pour les patrons d'extraction de notre outil d'annotation, miR Discovery, nous avons exploité un ensemble de connaissances biologiques concernant les miARN.

5.1 Constitution de ressources sur les miARN

Le point de départ de notre travail a d'abord été de mettre en exergue les règles concernant les miARN et les processus biologiques engagés lors de l'appariement des miARN et des ARNm. C'est un champ d'activité nouveau et prometteur sur lequel les biologistes veulent travailler et obtenir des résultats très rapidement pour être concurrentiels.

La table 1 présente un aperçu des règles surlignées par l'expert biologiste, et qui d'après lui contiennent l'information pertinente à extraire. Un type de règle en particulier représente un intérêt fondamental pour les biologistes, il s'agit des passages de textes qui traitent d'un lien éventuel entre une mutation sur un gène ou un miARN et une maladie. Nous retrouvons cette connaissance dans les règles 1 à 5 de la table 1. Notons que la connaissance exprimant un lien entre une mutation et/ou un gène et/ou un miARN sans allusion à aucune maladie est aussi une information potentiellement intéressante pour les biologistes. En effet, le biologiste peut être alerté sur ce lien existant et en vérifier si nécessaire la nature et les conséquences induites. Afin de pouvoir détecter dans les textes les termes relatifs aux mutations, aux gènes, aux miARN ainsi qu'aux maladies, il a été nécessaire d'identifier les terminologies respectives et de les charger dans l'outil d'extraction d'information.

Table 1. Liste des règles surlignées par l'expert biologiste, qui identifient clairement une connaissance d'intérêt pour les biologistes sur les miARN. On y retrouve également le PMID (identifiant unique des articles dans Pubmed) ainsi que les termes importants de la règle qui ont permis d'attirer l'attention du biologiste sur l'intérêt de la règle en question.

N°	Règle	PMID	Termes importants
1	Two novel mutations (C+7T/miR-16-1, G+19A/let-7e) have been shown to differentially modulate miRNA expression in vivo;	18778868	Mutations (C+7T/miR-16-1, G+19A/let-7e) / miRNA expression
2	however, only one (C+7T/miR-16-1) has been reported to be associated (P = 0.038) with a human disease: chronic lymphocytic leukemia (CLL)	18778868	(C+7T/miR-16-1) / chronic lymphocytic leukemia (CLL)
3	Among these, they identified one SNP in the 3' UTR of the cluster of differentiation 86 (CD86) gene, rs17281995 [C/G], that was significantly associated with colorectal cancer	18778868	SNP / CD86 / rs17281995 / colorectal cancer
4	Calin et al. also described at least two more novel, potentially CLL-associated miRNA mutations (G49T/miR-206 and +107A/miR-29b-2) that merit further experimental analysis.	18778868	CLL / associated miRNA mutations / (G49T/miR-206 and +107A/miR-29b-2)
5	The authors observed that the strongest association (P = 0.0001) was with rs12720208 [C/T], a SNP that the authors demonstrate mediates allele-specific in vitro targeting of miR-433 to the FGF20 3' UTR.	18778868	rs12720208 / targeting of miR-433 to the FGF20 3' UTR
6	As the target sites were designed to allow optimal 3' pairing, we conclude that G:U base-pairs in the seed region are always detrimental	15723116	G:U base-pairs / seed region are always detrimental
7	Surprisingly, a single 8mer seed (miRNA positions 1–8) was sufficient to confer strong regulation by the miRNA	15723116	single 8mer seed / miRNA positions 1–8 / strong regulation

Table 2. Terminologies ou bases de données utilisées pour construire les dictionnaires relatifs aux gènes, aux miARN, aux maladies et aux mutations.

Termes	**Source**	**Exemple**	**Nombre de termes récupérés**
Gènes	HUGO[7] SwissProt Tarbase	CD86	119 408
miARN	miRBase	let-7 mir-318	904
Maladies	Sous ensemble de la SNOMED	chronic lymphocytic leukemia (CLL)	4 443
Mutations	Construction manuelle par des graphes (Cf. figure 2)	rs17281995 G49T/miR-206	Infini

Le corpus utilisé a été collecté manuellement en envoyant la requête suivante à PubMed : *SNPs [MH] AND miRNAs [MH] AND human [MH].*

[MH] indique que le terme à gauche est un terme MeSH[8]: le Medical Subject Headings est un thésaurus biomédical proposant 25 186 termes en 2009 pouvant être utilisés pour décrire très précisément le contenu d'un document médical.

[7] http://www.hugo-international.org/

[8] http://www.nlm.nih.gov/mesh/

Le résultat de cette requête correspondait à 35 articles[9], parmi lesquels uniquement 21 étaient gratuitement disponibles en entier.

Le corpus exploité comporte donc 21 articles provenant de journaux différents, équivalents à 533 853 tokens, et d'une taille de 2,2 Mo.

5.2 Extraction de la connaissance relative aux miARN

Nous avons réuni toutes les phrases dans lesquelles nous retrouvons une proposition décrivant un lien entre une mutation, un gène, un miARN et/ou une maladie parmi les règles concernant les miARN surlignées par l'expert biologiste. Il est alors apparu qu'aucune phrase n'a la même structure syntaxique que les autres. En effet, il existe de très nombreuses manières d'exprimer ce lien biologique. Nous avons également noté que les phrases dans lesquelles nous retrouvons une mutation, un gène, un miARN et/ou une maladie sont forcément des phrases qui relatent un lien biologique entre ces différents éléments. Partant de cette hypothèse, la construction de patrons lexico-syntaxiques avec différents chemins syntaxiques n'est pas nécessaire, c.-à-d. que des patrons uniquement lexicaux de tri-occurrences (mutation et (gène ou miARN) et maladie) ou de quadri-occurrences (mutation et gène et miARN et maladie) sont suffisants pour extraire les connaissances qui intéressent les biologistes.

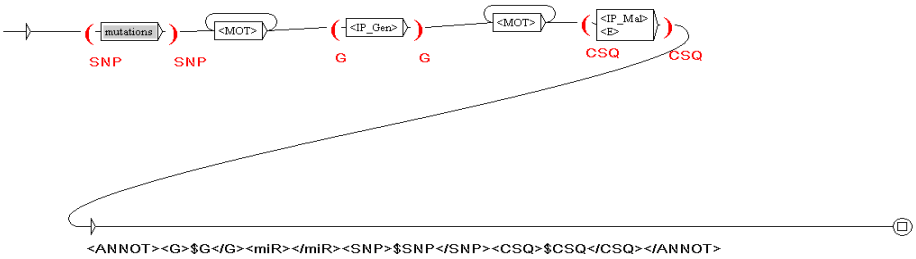

Fig. 3 – Patron de tri-occurrence reconnaissant une mutation suivi d'une succession de motS (<MOT>*), puis un nom de gène (<IP_Gen>) suivi d'une succession de mots (<MOT>*) et enfin une maladie (<IP_Mal>) ou rien (<E>). Si le patron reconnaît une phrase dans le texte, les termes déterminants pour l'extraction sont récupérés grâce aux variables définies : SNP pour les mutations, G pour le gène et CSQ pour la maladie. Le résultat de l'extraction sera donc :<ANNOT><G>$G</G><miR></miR><SNP>$SNP</SNP><CSQ>$CSQ</CSQ></ANNOT>

Les patrons sont donc construits pour détecter une tri-occurrence ou quadri-occurrence dans le texte.

La figure 3 montre un graphe représentant un patron de tri-occurrence pour détecter les phrases exprimant un lien entre une mutation, un gène et/ou une maladie. Il fait appel à un sous-graphe nommé mutations (boîte grisée) présenté à la figure 4. Il a été

[9] Accès à Pubmed le 25 novembre 2008

nécessaire de construire un patron qui reconnaisse une infinité de mutations différentes (Figure 4), car il n'existe pas de base de données répertoriant de manière exhaustive toutes les mutations existantes chez l'humain. Ce patron passe néanmoins à côté de la détection de plusieurs mutations car il existe dans la littérature de très nombreuses façons de les exprimer, variant presque d'un auteur à un autre, et ce, malgré une nomenclature bien définie[10]. La figure 4 illustre quelques façons d'exprimer les mutations.

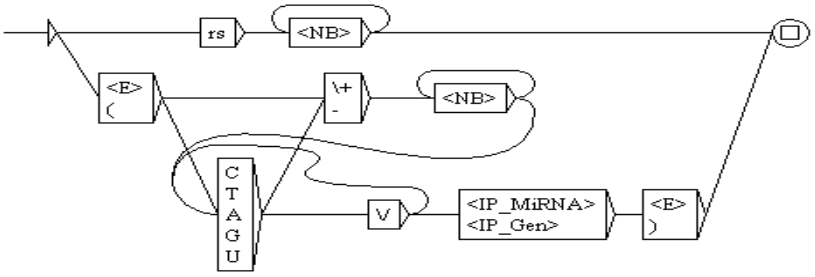

Fig. 4 – Patron de reconnaissance des mutations: il reconnaît des mutations de type : *rs* suivi d'un nombre (<NB>*), mais aussi de type *(G+19A/let-7e)* où *let-7e* est un miARN (<IP_MiRNA>), ou encore *+107C/miR-29b-6.*

5.3 Peuplement automatisé de l'ontologie des miARN

L'Unstructured Information Management Architecture (UIMA)[11] est une infrastructure logicielle initiée par le centre de recherche Alphaworks d'IBM et à présent repris par l'incubateur d'Apache. Elle fournit les bases pour développer un processus d'annotation sémantique, bien qu'elle ne donne pas pour autant des conseils sur la programmation et l'ordre des étapes de ce processus . Le Content Augmentation (CA) Manager développé dans le cadre du projet européen TAO[12] propose sur la base d'UIMA une liste d'étapes logiques, dont certaines sont optionnelles, qui vont être chaînées ensemble, enrichissant au fur et à mesure un schéma d'annotation pré-défini (Figure 5). Ces étapes peuvent être groupées en trois thématiques ou composants principaux : 1) *Extraire* la connaissance pertinente et annoter le contenu; 2) *Consolider* les résultats vis à vis du modèle de l'ontologie et du référentiel sémantique; 3) Sérialiser le schéma d'annotation dans divers formats et le *stocker* dans le référentiel sémantique.

Avant même d'initier la démarche d'annotation sémantique et de peuplement d'ontologie, il nous faut charger l'ontologie des miARN dans un référentiel sémantique comme Sesame[13] ou ITM[14]. Puis, la première étape d'extraction consiste à appeler l'outil

[10] http://www.genomic.unimelb.edu.au/mdi/mutnomen/recs.html
[11] http://www.research.ibm.com/UIMA/
[12] http://www.tao-project.eu
[13] http://www.openrdf.org/
[14] http://mondeca.com/index.php/en/intelligent_topic_manager

d'analyse linguistique décrit ci-dessus, miR Discovery, afin d'annoter automatiquement les articles biologiques de PubMed. Chaque nouvelle annotation ou instance générée (notamment de relations entre les miARN, mutations, gènes et maladies) est contrôlée dans le référentiel sémantique afin de vérifier la non-redondance de la connaissance et la préservation de la cohérence et de la qualité de la base de connaissances. Pour ce faire, un ensemble de règles de consolidations (Amardeilh, 2008) sont définies et appliquées sur chaque nouvelle annotation, instance de classe ou instance de propriété. Ces règles peuvent être simples comme vérifier qu'une instance de même classe et même libellé n'existe pas déjà dans la base de connaissances ou bien beaucoup plus complexes. Dans ce cas, les règles de consolidation peuvent être complétées par des règles de raisonnement et d'inférence. Les informations non valides vis à vis du modèle ontologique ou de la base de connaissances seront mises de côté avec le statut « à valider » afin que si l'application finale repose sur une interface de validation manuelle de ces annotations et instances, le CA Manager soit en mesure de remonter ces informations non valides pour permettre à l'utilisateur de les corriger et les désambiguïser si besoin.

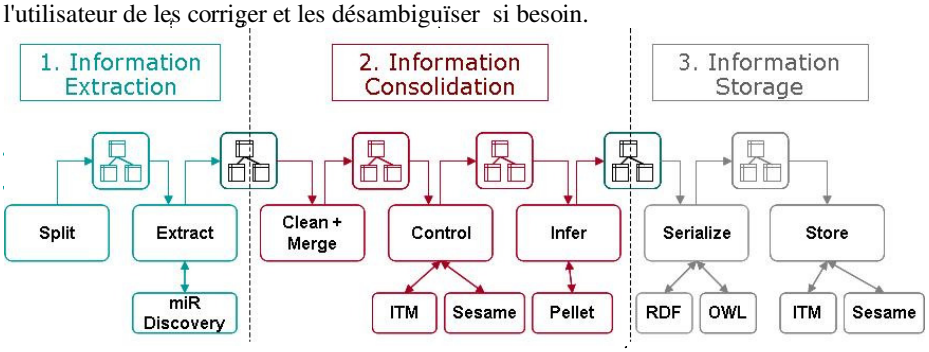

Fig. 5 - Un workflow modulaire basé sur UIMA

5.4 Evaluation et Résultats

Nous avons pu extraire des connaissances concernant les mutations ainsi que des indications de leur emplacement (miARN ou gène) mais aussi parfois leur lien éventuel avec une maladie. Ces connaissances sont extraites et proposées au format XML, comme l'indique la figure 6. Nous avons extrait 35 annotations différentes, c.-à-d. distinctes les unes par rapport aux autres (sur le modèle de la figure 6). Sur la totalité du corpus, 30 annotations provenant de 49 phrases différentes ont été réalisées manuellement par l'expert biologiste. Elles sont utilisées comme référence pour notre évaluation. Sur ces 49 phrases, certaines proposent donc des annotations redondantes.

Notre outil a permis de faire 35 annotations. Sur ces 35 annotations, 25 sont correctes mais certaines sont incomplètes (on ne retrouve pas la maladie par exemple) ou redondantes et, parmi elles, 15 sont strictement identiques aux annotations de référence faites par l'expert biologiste. Nos mesures de rappel et de précision sont données ci-après :

- Précision = 25 / 35 = 0,72
- Rappel = 15 / 30 = 0,50

Le chiffre du rappel est relativement bas car nous n'avons pas pris en compte les variantes morphologiques des maladies par exemple. En effet, notre outil ne détecte pas une phrase qui contient *« lung cancers »* car notre dictionnaire n'inclut que les formes au singulier *« lung cancer »*. Le chiffre de la précision souffre de la synonymie des noms de gènes avec des noms, des prénoms, ou des acronymes utilisés pour référencer des techniques en biologie. Ce problème n'est pas nouveau, et de nombreuses équipes y travaillent.

```
<ANNOT>
  <G>FGF20</G>
  <miR>miR-433</miR>
  <SNP>rs12720208</SNP>
  <CSQ />
  <PMID>18252210</PMID>
  <SENT>miR-433 Inhibits FGF20 Translation at SNP rs12720208 .</SENT>
</ANNOT>
```

Fig. 6 – Exemple de connaissances extraites au format XML, les balises <G> sont pour les gènes, <miR> pour les miARN, <SNP> pour les mutations, <CSQ> pour les maladies et <SENT> pour les phrases dont proviennent les annotations.

6 Conclusion

Nous venons de présenter dans cet article la notion de filtrage sémantique à travers les activités d'annotations sémantiques et de peuplement d'ontologie traitant des miARN. Nous avons vu que ce filtrage est intrinsèquement lié à la modélisation d'une ontologie de domaine dans le cadre du Web Sémantique. En effet, cette ontologie va représenter les concepts, attributs et relations d'un domaine à l'aide d'un langage de représentation des connaissances orienté Web comme OWL. Elle sera instanciée à partir des extractions linguistiques, par exemple les interactions entre des miARN et ARNm identifiés dans les articles de PubMed. L'ontologie que nous avons modélisée sert de médiateur et d'accès aux différentes terminologies de l'application. Celle-ci repose sur une architecture SOA (Service Oriented Architecture) s'appuyant sur les web services proposés par le CA Manager. Plus ce modèle de représentation de la connaissance dans lequel seront exprimées les instances est formel, plus les services proposés seront « intelligents ». Les biologistes pourront interroger les connaissances en fonction de différents points de vue : recherche par mots-clefs (annotations sémantiques), recherche multicritère (en fonction des concepts, relations et attribut du modèle), ou encore recherche par extension sémantique (terminologies). Les agents logiciels pourront également raisonner et inférer de la nouvelle connaissance et ainsi dégager un sens implicite contenu dans le document d'origine (Laublet, 2007). Le peuplement d'ontologie à l'aide des techniques du Web sémantique ouvre donc des perspectives intéressantes à de nombreuses applications comme la recherche d'informations sémantiques, la catégorisation, la composition de documents, etc. L'extraction de connaissances donne un rappel de 50%, et la précision des résultats est de 72%. Ces deux mesures peuvent être améliorées et nous y travaillons.

Remerciements

Cette étude a été financée par le projet Microbio (programme Stic-Amsud).

Références

Amardeilh, F. (2008). Semantic Annotation & Ontology Population. *Semantic Web Engineering in the Knowledge Society, ISI Global*.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet, 25*(1), 25-29.

Bourigault, T. D., Aussenac-gilles, N., & Charlet, J. (2004). Construction de ressources terminologiques ou ontologiques à partir de textes: un cadre unificateur pour trois études de cas. *Revue d'Intelligence Artificielle, 18,* 87-110.

Brisson, L., & Collard, M. (2007). *Intérêt des systèmes d'information dirigés par des ontologies pour la fouille de données*. Rapport de recherche Projet Execo, Nice, Sophia Antipolis: CNRS.

Eilbeck, K., Lewis, S. E., Mungall, C. J., Yandell, M., Stein, L., Durbin, R., et al. (2005). The Sequence Ontology: A tool for the unification of genome annotations. *Genome Biology, 6*(5).

Griffiths-Jones, S. (2004). The microRNA Registry. *Nucleic Acids Research, 32*(Database Issue), D109-D111.

Hearst, M. A. (1992). *Automatic acquisition of hyponyms from large text corpora.* Paper presented at the International Conference on Computational Linguistics (COLING'92), Nantes.

Hignette, G. (2007). *Annotation sémantique floue de tableaux guidée par une ontologie, Thèse de doctorat,* AgroParistech.

Jilani, I., Grabar, N., & Jaulent, M.-C. (2006). *Fitting the finite-state automata platform for mining gene functions from biological scientific literature.* Paper presented at the Semantic Mining in Biomedicine, Jena (Germany).

Laublet, P. (2007). Web sémantique et ontologies. In Hermès (Ed.), *Humanités numériques Nouvelles technologies cognitives et concepts des sciences sociales* (Vol. 1). Paris.

Morin, E. (1999). Acquisition de patrons lexicosyntaxiques caractéristiques d'une relation sémantique. *Traitement Automatique des Langues (TAL), 40(1)*, 143-166.

Prié, Y., & Garlatti, S. (2004). Méta-données et annotations dans le Web sémantique. *Revue I3 Information - Interaction - Intelligence, 4*, 45-68.

Sethupathy, P., Corda, B., & Hatzigeorgiou, A. G. (2006). TarBase: A comprehensive database of experimentally supported animal microRNA targets. *RNA, 12*, 192-197.