Apprendre de la littérature scientifique : Les réseaux de signalisation en biologie systémique

Flavie Landomiel[1], Cathy Guérineau[2,] Anubhav Gupta[2,]
Denis Maurel[2], Anne Poupon[1]
(1) PRC, INRA/CNRS/Université de Tours
(2) Université de Tours, Lifat
Denis.maurel@univ-tours.fr, anne.poupon@inra.fr

RÉSUMÉ

Cet article a pour but de montrer la faisabilité d'un système de fouille de texte pour alimenter un moteur d'inférences capable de construire, à partir de prédicats extraits des articles scientifiques, un réseau de signalisation en biologie systémique. Cette fouille se réalise en deux étapes : la recherche de phrases d'intérêt dans un grand corpus scientifique, puis la construction automatique de prédicats. Ces deux étapes utilisent un système de cascades de transducteurs.

ABSTRACT

Literature-based discovery: Signaling Systems in Systemic Biology.
This paper aims to prove the feasibility of a text mining system to provide an inference engine that builds a signaling system in Systemic Biology. This engine uses predicates extracted from scientific papers. Our text mining system proceeds in two steps. First, it searches interesting sentences in a big scientific corpus. Second, it automatically builds predicates from these sentences. These two steps use finite state transducer cascades.

MOTS-CLÉS : Fouille de texte ; Réseaux de signalisation ; Biologie systémique ; Cascade de transducteur ; Cassys ; Unitex, Prédicats.

KEYWORDS: Text mining; Signaling system; Systemic Biology; Finite state transducer cascades; Cassys; Unitex, Predicates.

1 Introduction

Aujourd'hui, en biologie, la profusion incroyable d'articles scientifiques rend impossible leur suivi par un chercheur. Par exemple, le mot-clé *Signaling* sur Pubmed retourne 480 publications juste pour le mois de septembre 2017. Impossible aussi de "suivre l'actualité" concernant des dizaines de milliers de gènes ou protéines. Des outils automatiques sont nécessaires pour cela.

Le traitement automatique des langues et la fouille de texte sont largement utilisés dans le domaine clinique (Demner-Fushman et Elhadad, 2016 ; Meystre et al., 2008) et biomédical (Zweigenbaum et al., 2007) avec de nombreuses campagnes d'évaluation (Huang et Lu, 2015). Il s'agit pour les chercheurs de retrouver des articles parmi des milliers d'autres. Mais aussi, de découvrir de

nouveaux résultats scientifiques en mettant "bout-à-bout" des résultats disséminés, la *Literature-based discovery* (Weeber et al., 2005). Notre projet vise la détection des interactions entre protéines et/ou gènes. Plusieurs travaux ont déjà été effectués dans ce sens, principalement en étudiant les co-occurrences entre protéines dans les articles scientifiques (Franceschini et al., 2013), la recherche s'effectuant sur des résumés (Hoffmann et Valencia, 2004) ou sur le texte complet (Rzhetsky et al., 2004 ; Miwa et al., 2010 ; Bunescu et al., 2006). Cependant seulement 30% des paires de protéines détectées sont réellement en interaction. Notre projet se rapproche plus des campagnes d'évaluation BioCreative, particulièrement la tâche 5[1] ou BioNLP-ST[2].

Cependant, il en diffère fortement pour la raison suivante : nous avons choisi pour notre part de cibler uniquement la partie factuelle des articles biologiques, à savoir les résultats des expériences réalisées. Ceux-ci, dans un article de biologie ou de médecine, se trouvent dans la partie "Résultats" où sont décrites précisément les expériences. Nous évitons en particulier les introductions et les discussions, tout comme les interprétations des auteurs, ainsi que les citations ou références à des travaux antérieurs. Notre but ultime est la construction automatique de réseaux de signalisation en biologie systémique. Voici notre schéma d'action : créer un corpus d'articles scientifiques à partir de trois mots-clés (deux protéines et un prédicat) ; extraire de ce corpus ce que nous considérons comme des *phrases d'intérêt* ; transformer ces phrases en relations *prédicat-arguments* ; et, enfin, utiliser un système expert pour construire un réseau de signalisation à partir des structures prédicatives mises en évidence. Ensuite, nos résultats seront testés par des collègues biologistes afin de prouver leur véracité scientifique.

La recherche des phrases d'intérêt et leur transformation en structure prédicative sera réalisée en utilisant un système de cascades de transducteurs (Abney, 1996 ; Friburger et Maurel, 2004) implanté dans le logiciel libre Unitex (Paumier, 2003) sous forme de cascades de graphes (Maurel et al., 2011). La construction du réseau de signalisation se fera par un système d'inférence automatique basé sur les travaux de (Gloaguen et al., 2011) et (Rougny et al., 2013). Ce système fonctionnait jusqu'à présent avec des prédicats écrits manuellement par des lecteurs humains et a déjà démontré son efficacité. Notre objectif ici est donc de prouver la faisabilité d'un système de fouille de texte pour alimenter ce moteur d'inférences. Dans l'état actuel du projet, la première partie est réalisée, la seconde dispose d'un premier prototype aux résultats encourageants et la troisième est à venir.

Cet article présentera donc la biologie systémique et les réseaux de signalisation (section 2), puis la recherche des phrases d'intérêt (section 3) et la construction des structures prédicatives (section 4), avant de conclure en présentant la suite du travail.

2 Biologie systémique et réseaux de signalisation

Les cellules développent des réponses spécifiques aux stimuli envoyés par l'organisme, le plus souvent par la mise en circulation d'hormones qui se lient à des récepteurs spécifiques à la surface des cellules. Cette liaison déclenche des cascades de réactions moléculaires, appelées transduction du signal. Nous nous focalisons sur la transduction du signal par les récepteurs couplés aux protéines G (RCPG), qui correspondent à plus de huit cents récepteurs différents. Ce sont des cibles pharmaceutiques idéales qui représentent aujourd'hui environ 40% des médicaments sur le marché. Or, seulement 15% des récepteurs sont « utilisés » par la pharmacopée. Les voies de signalisation

[1] http://www.biocreative.org/tasks/biocreative-vi/track-5/
[2] http://2016.bionlp-st.org/tasks/ge4

sont encore mal connues ; une meilleure connaissance permettra la mise au point de médicaments plus efficaces et ayant moins d'effets secondaires indésirables

La mise au point et les tests qui suivent ont été réalisés pour deux protéines (*ERK* et *arrestin*) et un prédicat (*phosphorylation*). Nous avons donc téléchargé en interrogeant sur ces trois mots-clés tous les articles scientifiques disponibles au format texte sur les bases Istex et NCBI[3]. Soit 3 255 documents. Comme il a été dit en introduction, nous nous focalisons sur la partie "Résultats" et nous avons donc éliminés les textes pour lesquels les trois mots clés utilisés ne se trouvaient pas tous les trois dans cette partie. Ce qui nous a permis de constituer un corpus de 1 282 documents (soit 40% des documents initialement téléchargés).

3 Recherche des phrases d'intérêt[4]

Nous appelons *phrase d'intérêt*, une phrase, de la partie "Résultats", contenant un groupe verbal qui fait référence à la démonstration d'un résultat scientifique impliquant deux protéines minimum. Pour la recherche des phrases d'intérêt, nous avons extrait, de notre corpus de 1 282 documents, un corpus de travail de cinq articles scientifiques et un corpus d'évaluation de vingt-sept articles pris au hasard parmi les 1 282 documents disponibles. Ce corpus représente des articles divers et variés en terme d'année de parution, de nationalité de l'auteur et de journal de publication.

3.1 Étude du corpus de travail

Nous illustrons ci-dessous notre concept de *phrases d'intérêt* par des exemples extraits d'un de nos cinq articles, celui de (Wang et al., 2005) :

1. We found that only phosphorylated ERK bound to Cdc25A.
2. These data provide evidence that the Cdc25A-ERK interaction can be independent of EGFR activation.
3. As shown in Figure 1B, GST-Cdc25A bound to ERK in vitro, whereas glutathione almost completely blocked this binding.

Ces phrases font directement référence à l'article lui-même : soit par un résultat d'expérience (*We found*) ; soit en résumant les diverses démonstrations citées dans le paragraphe *These data provide* ; soit en mentionnant la figure démontrant la phrase conclusive *As shown in Figure 1B*.

A l'inverse, nous voulons éviter de retrouver des phrases parasites. Parmi celles-ci, les phrases descriptives annonçant ce que les auteurs ont l'intention de montrer, par exemple :

4. We next examined whether Cpd 5-induced ERK phosphorylation can be independent of MEK, its direct up-stream kinase activator.

Mais aussi des phrases qui font le lien avec des résultats précédemment publiés ou avec de la bibliographie, telles que :

5. We previously reported that Cpd 5, a Cdc25A inhibitor, caused prolonged EGFR activation, which in turn triggered ERK phosphorylation and cell growth inhibition (Wang et al., 2000, 2002).

[3] https://www.istex.fr/ et https://www.ncbi.nlm.nih.gov/.
[4] Cette partie a fait l'objet d'une communication dans un atelier (Landomiel et al., 2017).

3.2 Les prétraitements

Une des fonctions principales d'Unitex concerne la création et l'application de dictionnaires spécifiques pour l'analyse de corpus. Dans le cas de notre projet, nous avons créé quatre dictionnaires pour prendre en compte la totalité des termes propices à notre analyse : tout d'abord un dictionnaire rassemblant les diverses techniques mises en place en laboratoire, ainsi que tout le lexique inhérent au domaine (issu de la bibliographie ainsi que des sites spécialisés) ; puis un dictionnaire de protéines (issu de la base de données *UniProt*) et deux dictionnaires comprenant les divers systèmes cellulaires et les composés chimiques usuels (issus des sites spécialisés et du NCBI).

Ce dictionnaire est fléchi et complété par deux graphes : le premier pour les formes conjuguées avec des auxiliaires, le second pour certaines formes polylexicales spécifiques aux anticorps. Ces deux graphes sont précédés d'un graphe de découpage en phrases, version anglaise inspirée de (Friburger et al., 2000).

3.3 La cascade

Suite à l'annotation manuelle qui nous a permis de mettre en lumière les constructions de phrases récurrentes dans les articles et la création des dictionnaires, nous avons créé trois autres graphes pour décrire les relations entre les protéines (Figure 1).

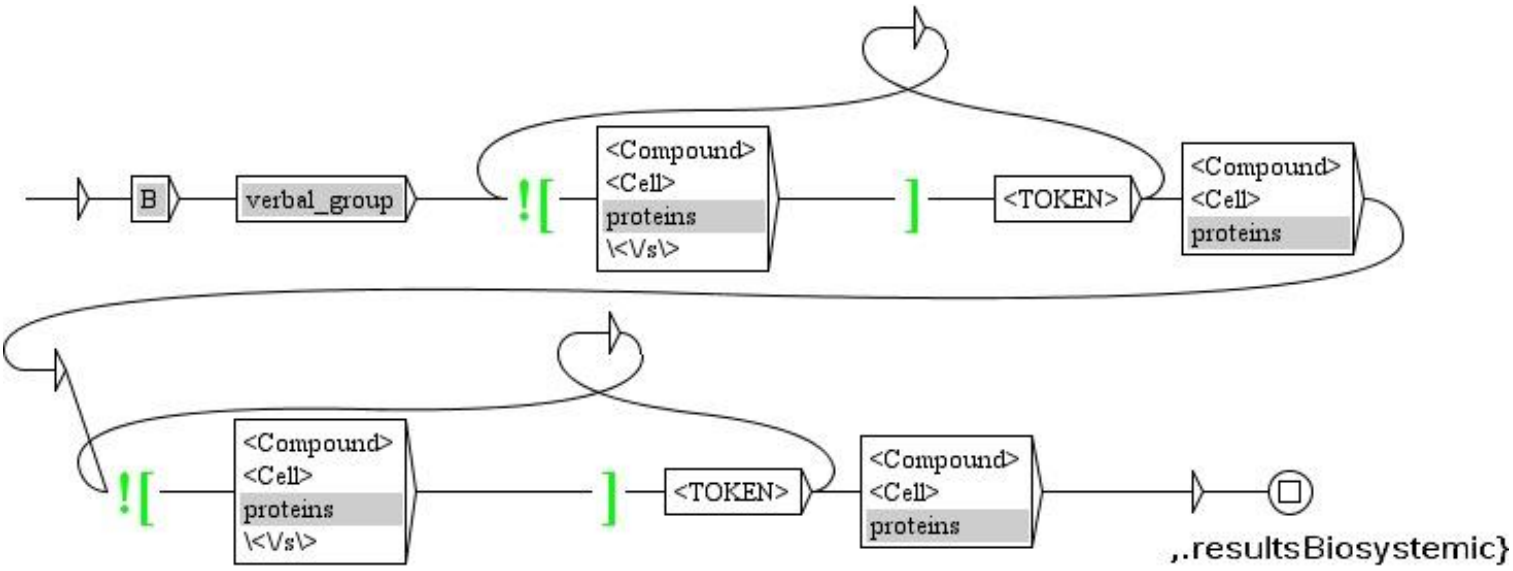

Figure 1 : Un des graphes désignant les relations entre les protéines

Notre cascade est subdivisée en deux sous-cascades. À la fin de la première sous-cascade, toutes les phrases d'intérêt sont identifiées en entier et le texte est balisé au format XML propre à CasSys. Dans la seconde sous-cascade, les balises extra-numéraires sont supprimées ainsi que les phrases qui pourraient être retrouvées par leur structure, mais dont le sens n'est pas recherché. Les verbes trop descriptifs tels que *was collected*, *was probed* mais aussi les phrases commençant par *We next*, *To further* ou encore *In order* qui ne vont pas aboutir à des phrases conclusives y sont intégrés. Enfin, les deux derniers graphes éliminent toutes les balises intermédiaires et crée un fichier rassemblant les phrases clés à la suite les unes des autres.

Cette cascade a été passée sur les 1 282 documents de notre corpus, ce qui nous donne un total de 62 655 phrases extraites.

3.4 Évaluation

Comme il a été dit, notre corpus d'évaluation compte vingt-sept articles, ce qui représente environ 14 000 phrases, dont 4 000 dans la partie résultats. Nous avons calculé les mesures classiques de rappel et de précision (Table 1)

Rappel	90%
Précision	81%

Table 1 : Les résultats de l'évaluation de la première cascade

Malgré des mesures de rappel et de précision très encourageantes, il est toujours possible d'améliorer ces valeurs. Par exemple, nous avons apporté une légère correction au graphe des fins de phrase, en ajoutant la possibilité d'un début de phrase par un chiffre suivi d'une minuscule, ce qui n'avait pas été prévu, mais peut correspondre au nom d'une protéine. D'autre part, les phrases commençant par *when* et celles contenant les mots *previously* et *et al.* avaient été éliminées lors de l'annotation et sont donc non reconnues par la cascade ; or, il s'avère qu'elles correspondent à environ un tiers des phrases manquantes. Il faudra expérimenter si leur réintégration peut augmenter le rappel sans impacter négativement la précision.

4 Construction des structures prédicatives

4.1 Les entités et prédicats traités

Le travail présenté ci-dessous nous a permis de valider notre démarche en testant la construction de trente-cinq entités et prédicats à travers un premier prototype. Quelques exemples sont donnés sur la Table 2. Ces entités et prédicats ont été choisis sur la base des travaux déjà réalisés, cités précédemment, à savoir (Gloaguen et al., 2011) et (Rougny et al., 2013). Ils sont proches de ceux définis dans les campagnes BioCreative et BioNLP-ST évoquées en introduction.

ENTITÉS et PRÉDICAT	DÉFINITION
cell(CL)	CL is a cell type
expressed(G,E,CL,MET)	Gene G is expressed (positive, hypothesis)/not expressed, in Cell type CL, using Method MET
molecule(X)	X is a molecule
particle(X)	X is a particle (molecule, bacteria, virus, etc...)
protein(P)	P is a protein
reactionModulation(X,Y,Z,E,CL,MET)	Signal X modulates the reaction Molecule Y -> Molecule Z, in Cell line CL (can be left void), using Method MET

Table 2 : Exemples de quelques entités et prédicats recherchés

4.2 La cascade

Pour construire les structures prédicatives, une seconde cascade est mise en place. Cette cascade travaille non plus sur les documents du corpus, mais sur les phrases d'intérêt qui en ont été extraites. Elle est, elle aussi, divisée en deux sous-cascades.

La première sous-cascade prépare le document en y repérant les informations qui seront nécessaires à la construction des entités et prédicats. En préliminaire, elle supprime les ellipses concernant les protéines ; ainsi *α- and β-arrestin* devient *α-arrestin and β-arrestin*. Les différentes constructions verbales sont identifiées aussi, l'actif et le passif, mais aussi, c'est important, la négation. Le graphe suivant mets entre balises les métadonnées comme par exemple l'abréviation du mot *Fig* très fréquemment utilisé pour *figure*, mais qui est également le nom d'une protéine. Sans ce graphe, des formules telles que *Fig. 1A* seraient balisées *<Protein>Fig</Protein>. <Gene>1A</Gene>*. Les données numériques résultant des expériences décrites dans les articles sont également balisées afin d'éviter leur intégration dans un nom de gènes. Les graphes suivants repèrent les noms de gènes, protéines ou molécules qui apparaissent dans ces phrases. Puis sont balisées les méthodes utilisées et les prédicats. Voir par exemple la Figure 2.

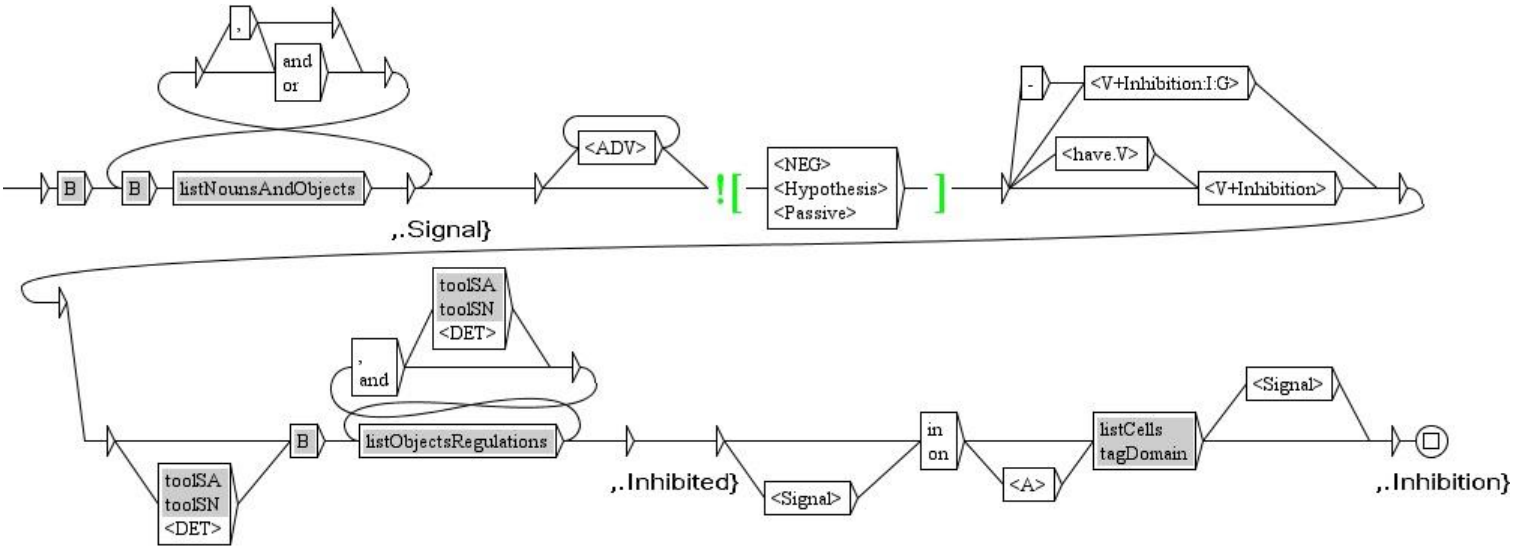

Figure 2 : Le graphe correspondant à un signal d'inhibition

Enfin, la seconde sous-cascade transforme le texte balisé par la première en des structures prédicat-arguments.

La Table 3 présente l'exemple complet d'une séquence issue d'une phrase d'intérêt.

4.3 Évaluation

Pour l'évaluation, nous avons créé un corpus constitué des phrases d'intérêt extraites de cinq documents pris au hasard parmi les 1 282 documents disponibles. Ce corpus représente un total de 226 phrases d'intérêt. Il y a en moyenne 45 phrases d'intérêt par articles.

Comme précédemment, nous avons calculé le rappel et la précision (Table 4). Cependant, les entités étant plus faciles à repérer que les prédicats, nous avons recalculé le rappel et la précision uniquement pour ceux-ci. Le rappel diminue, mais la précision augmente légèrement.

Au final, les résultats sont globalement corrects, sachant que, d'une part, ce prototype sera amélioré et, d'autre part, le système d'inférences qui suit pourra rejeter certains prédicats.

Texte brut	a robust activation of ERKs in δ-OR-expressing HEK-293 cells
Texte balisé	a robust <Activation>activation of <Activated type="Protein">ERKs</Activated> in <Reaction type="Expression"><Cell><Protein>δ-OR</Protein>-expressing <Cell>HEK-293 cells</Cell></Cell></Reaction></Activation>
Entités et prédicats	molecule(ERKs) particle(ERKs) protein(ERKs) cell(HEK-293) molecule(δ-OR) particle(δ-OR) protein(δ-OR) expressed(δ-OR, HEK-293, positive) reactionModulation(unknown signal, ERKs_inactive, ERKs_active, increase, δOR-expressing HEK-293)

Table 3 : Traitement d'une séquence issue d'une phrase d'intérêt

	Entités et prédicats	Prédicats
Rappel	88%	65%
Précision	76%	79%
F-mesure	82%	71%

Table 4 : Les résultats de l'évaluation de la seconde cascade

5 Conclusion et perspectives

Nous venons de montrer la faisabilité d'un système de fouille de texte pour alimenter un moteur d'inférences capable de construire, à partir de prédicats extraits des articles scientifiques, un réseau de signalisation en biologie systémique.

Ceci en deux étapes : la recherche de phrases d'intérêt dans un grand corpus scientifique, puis la construction automatique de prédicats. Ces deux étapes utilisent un système de cascades de transducteurs.

Dans nos perspectives, nous comptons rendre opérationnelle la seconde cascade, afin d'obtenir de meilleurs résultats étendus à une liste plus importantes de prédicats. Puis de créer un enchaînement complet, nos deux cascades et le moteur d'inférence, qui pourra, à partir de la donnée de deux protéines et d'un prédicat, construire des réseaux de signalisation.

Remerciements

Ce projet a été financé par le programme "Chantiers d'avenir" instauré dans le cadre du projet Istex (ANR-10-IDEX-0004-02).

Références

ABNEY S. (1996), Partial Parsing via Finite-State Cascades, *Workshop on Robust Parsing*, 8th *European Summer School in Logic, Language and Information*, Prague, Tchèquie, 8-15.

BUNESCU R., MOONEY R., RAMANI A., MARCOTTE E. (2006). Integrating Co-occurrence Statistics with Information Extraction for Robust Retrieval of Protein Interactions from Medline, in Proceedings of the workshop on linking natural language processing and biology: towards deeper biological literature analysis. 49–56.

DEMNER-FUSHMAN D, ELHADAD N. (2016). Aspiring to Unintended Consequences of Natural Language Processing: A Review of Recent Developments in Clinical and Consumer-Generated Text Processing. Yearb Med. Inform. 224-233.

FRANCESCHINI A., SZKLARCZYK D., FRANKILD S., KUHN M., SIMONOVIC M., ROTH A., LIN J., MINGUEZ P., BORK P., VON MERING C., JENSEN L. J. (2013). STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res. 808-815.

FRIBURGER N., DISTER A., MAUREL D. (2000), Améliorer le découpage des phrases sous Intex, *Revue Informatique et Statistique dans les Sciences Humaines*, vol. 36, n°1-4, p. 181-200.

FRIBURGER N., MAUREL D. (2004), Finite-state transducer cascade to extract named entities in texts, *Theoretical Computer Science*, vol. 313, 94-104.

GLOAGUEN P., CRÉPIEUX P., HEITZLER D., POUPON A., REITER E. (2011). Mapping the follicle-stimulating hormone-induced signaling networks. Front Endocrinol. 2:45

HOFFMANN R., VALENCIA A. (2004). A gene network for navigating the literature. Nature Genetics 36, 664.

HUANG C.C., LU Z. (2015). Community challenges in biomedical text mining over 10 years: success, failure and the future. Brief Bioinform. 17(1):132-44.

LANDOMIEL F., GUPTA A., MAUREL D., POUPON A. (2017). Préliminaire à la construction d'un réseau de signalisation en biologie systémique. Atelier *Fouille de Textes - Text Mine*, en conjonction avec *EGC 2017*, Grenoble, 24 janvier[5].

MAUREL D., FRIBURGER N., ANTOINE J.-Y., ESHKOL-TARAVELLA I., NOUVEL D. (2011). Cascades de transducteurs autour de la reconnaissance des entités nommées. *Traitement automatique des langues*, 52(1):69-96.

MEYSTRE S. M., SAVOVA G. K., KIPPER-SCHULER K. C., HURDLE J. F. (2008). Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med. Inform. 128-144.

[5] http://vincentlemaire-labs.fr/TM2017/Actes_TextMine17.pdf

MIWA M., SÆTRE R., KIM J.-D., TSUJII J. (2010). Event extraction with complex event classification using rich features. J. Bioinform. Comput. Biol. 08, 131–146.

PAUMIER S. (2003), *De la Reconnaissance de Formes Linguistiques à l'Analyse Syntaxique*, Thèse de Doctorat en Informatique, Université de Marne-la-Vallée.

ROUGNY A., FROIDEVAUX C., YAMAMOTO Y., INOUE K. (2013). Translating the SBGN-AF language into logic to analyze signalling networks. LNMR 2013 (First International Workshop on Learning and Nonmonotonic Reasoning). 44-55.

RZHETSKY A., IOSSIFOV I., KOIKE T., KRAUTHAMMER M., KRA P., MORRIS M., YU H., DUBOUÉ P. A., WENG W., WILBUR W. J., HATZIVASSILOGLOU V., FRIEDMAN C. (2004). GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. J. Biomed. Inform. 37, 43–53.

WANG Z., ZHANG B., WANG M., CARR B. I. (2005). Cdc25A and ERK interaction: EGFR independent ERK activation by a protein phosphatase Cdc25A inhibitor, Compound 5. *Journal of Cellular Physiology* 204(2), 437–444.

WEEBER M., KORS J. A., MONS B. (2005). Online tools to support literature-based discovery in the life sciences. Brief. Bioinform. 6, 277–286.

ZWEIGENBAUM P., DEMNER-FUSHMAN D., YU H., COHEN K. B. (2007). Frontiers of biomedical text mining: current progress. Brief Bioinform. 8(5):358-75.