Extraction d'informations sur la régulation transcriptionnelle de gènes à partir d'articles biomédicaux 2008

Julien Lorec

UNIVERSITÉ DE NANTES
Faculté de Médecine

Extraction d'informations sur la régulation transcriptionnelle de gènes à partir d'articles biomédicaux

Thèse de doctorat

École doctorale : Chimie Biologie
Discipline : Sciences de la Vie et de la Santé
Spécialité : Bioinformatique

Présentée et soutenue publiquement par

LOREC Julien

le ff mm yyy, devant le jury composé de

Président	M. HOULGATTE Rémi, INSERM UMR915, Nantes
Rapporteurs	XXX
	M. MALTHIÈRY Yves, INSERM U694/Université d'Angers
Examinateurs	XXX
	M. RAMSTEIN Gérard, LINA-CNRS/École Polytechnique de l'Université de Nantes
Directeur de thèse	M. JACQUES Yannick, INSERM U601, Nantes

A mon frère.

Remerciements

Je remercie Yannick Jacques et Gérard Ramstein pour m'avoir dirigé pendant toute cette thèse et pour leur infinie patience. Merci.

Je remercie Rémi Houlgatte d'avoir accepté d'être le président de mon jury de thèse.

Je remercie XXX et Yves Malthièry d'avoir accepté d'être les rapporteurs de ce travail.

Je remercie XXX d'avoir accepté de participer à mon jury de thèse.

Je remercie Pascale Kuntz, Henri Briand, Fabrice Guillet, Julien Blanchard pour leur accueil, leur soutien, leur compétence et leur gentillesse.

Je remercie aussi tous les membres de l'équipe 3 de l'U601 pour leurs nombreux conseils et leur expertise.

Merci à mes (anciens) collègues de bureau et (toujours) amis. Un grand merci à Jérôme Mikolajczak, à Nicolas Beaume et à Jérôme David.

Merci aussi à tous les ex-doctorants et doctorants de l'U601 pour leur sympathie.

Résumé

FR

La cartographie des réseaux de régulation de la transcription des gènes et des mécanismes moléculaires impliqués sont des problématiques importantes pour les biologistes. Les ressources bibliographiques de biologie moléculaire sont une mine prodigieuse d'informations expérimentales qui couvrent l'état de l'art actuel dans le domaine de l'expression de gènes. Cependant, en raison de la taille gigantesque que représentent les données textuelles du domaine, des méthodes automatisées doivent être mises au point afin d'explorer ces données de manière systématique.

Dans cette thèse, nous proposons un ensemble de méthodes pour fouiller la littérature de biologie moléculaire et extraire les faits pertinents en relation avec l'expression de gènes humains. Nous présentons tout d'abord une procédure générique destinée à l'extraction d'entités nommées candidates à partir des textes. Celle-ci combine une approche d'identification à base de règles de groupes nominaux en tant qu'entités nommées candidates avec une étape de mise en correspondance au sein de dictionnaires expertisés et élaborés à partir de ressources terminologiques publiques. Des techniques de désambigüisation spécifiques au domaine sont aussi présentées afin de déterminer la nature réelle de l'entité nommée identifiée. Nous détaillons ensuite une méthode qui permet à la fois de d'extraire les relations pertinentes établies entre les entités nommées et de retrouver certaines caractéristiques de ces associations grâce à une analyse syntaxique dite profonde et l'utilisation de structures prédicat-arguments. Nous montrons que l'acquisition de la sémantique à partir de la syntaxe peut être séparée en deux phases distinctes afin de réduire le cout associé à la conception manuelle de règles d'extraction spécifiques au domaine. Finalement les performances du système sont évaluées à l'aide d'un corpus annoté de publications complètes de biologie moléculaire. Les résultats sont prometteurs et malgré la nature hétérogène des données extraites, le système présente des performances à la fois homogènes et compatibles avec la montée en charge.

EN

Charting transcriptionally regulated networks of genes and gathering related molecular mechanisms are important issues for biologists. The molecular biology literature is a very rich mine of experimental information that encompasses the current state of knowledge in the gene expression domain. However, due to its tremendous size, automated methods must be devised in order to explore these data in a systemic way.

In this thesis, we propose a method set for mining the molecular biology literature and extracting relevant facts about human gene expression regulation. We first present a generic methodology to extract potential named entities from texts. This combines rule-based identification of noun phrases as candidate named entities with matching against manually cleaned dictionaries from public sources. Domain-specific disambiguation techniques are also reported in order to help classifying the true nature of an identified named entity. Then we detail a procedure for both retrieving relevant relationships between named entities and their associated features using a deep syntactic analysis and predicate-argument structures. We show that the acquisition of semantics from syntax can be split into several distinct phases so as to lessen the labour usually associated with the design of domain-specific extraction rules. Finally the performance of the system is evaluated using an annotated corpus of specialized full-text publications. The results are promising and despite the heterogeneous nature of the information to retrieve from the data set, the system exhibits homogeneous and highly-scalable performances.

Table des matières

Table des figures

Liste des tableaux

Liste des abréviations

Pour des raisons de lisibilité, la signification d'une abréviation ou d'un acronyme n'est souvent rappelée qu'à sa première apparition dans le texte d'un chapitre. Les noms propres sont précisés en italiques.

ADN	Acide désoxyribonucléique
ARN	Acide ribonucléique
DC	Théorie de Dépendance Conceptuelle
DNA	= **ADN**
EEN	Extraction des Entités Nommées
EI	Extraction d'Information
EN	Entité nommée
FdT	Fouille de Texte
GM	Granulocyte Macrophage
GO	*Gene Ontology*
HMM	Modèles de Markov cachés
HUGO	*Human Gene Nomenclature Database*
IA	Intelligence Artificielle
IEN	Identification des Entités Nommées
IL	Interleukine
LGP	*Link Grammar Parser*
MCP	*Macrophage Chemoattractant Protein*
MUC	*Message Understanding Conferences*
NFKB	*Nuclear Factor Kappa B*
POS	Part Of Speech (ou fonction grammaticale)
REN	Reconnaissance des Entités Nommées
RI	Recherche d'Information
RKIP	*RAF Kinase Inhibitor Protein*
RNA	= **ARN**
SVM	Machines à vecteurs de support
TALN	Traitement Automatique de la Langue Naturelle
TFD	*Transcription Factors Database*
TFIID	*Transcription factor IID*
TRRDSITE	*Transcription Regulatory Regions Database Site*

Introduction

Les années 1990 ont été marquées par une croissance sans précédent à la fois dans la production de données biomédicales et dans la quantité de publications scientifiques qui en discutent. Un des buts ultimes de la biologie moderne est de réussir à traduire à grande échelle toutes les données accumulées en connaissance sur les processus biologiques complexes et d'obtenir des modèles précis du fonctionnement de la cellule vivante et des organismes.

L'ensemble des informations concernant les entités biologiques telles que les gènes, les protéines, les maladies, les médicaments et leurs rôles au sein des processus biologiques est mis à disposition dans la littérature scientifique. Les récentes avancées des technologies génomiques et protéomiques ont eu pour conséquence de générer une somme phénoménale d'informations en rapport avec les gènes et les protéines notamment. Cette abondance d'information biologique et le nombre de documents scientifiques qui en discutent ont pour contrepartie de rendre difficile leur accès et de complexifier leur interprétation.

Ainsi, les bases de données biologiques, parce qu'elles permettent de centraliser et de structurer toute cette information, sont devenues un complément essentiel aux ressources bibliographiques de biologie. Au fil des années, de plus en plus des données issues de la littérature seront répertoriées dans ces bases de données. Néanmoins, les questions liées aux modalités de la mise à jour de leur contenu restent ouvertes. Bien que de nombreux résultats expérimentaux soient déjà mis à disposition dans les bases de données biologiques, ce travail est en très grande majorité réalisé manuellement par des experts du domaine. L'augmentation toujours croissante du nombre de nouvelles publications biologiques depuis ces vingt dernières années a pour conséquence de précariser cette situation. L'extraction manuelle des informations contenues dans les textes est devenue aujourd'hui presque impossible et extrêmement couteuse.

Afin de répondre à cette problématique, des méthodes automatisées doivent être désormais mises en œuvre. Depuis quelques années déjà, les travaux appliqués à la biologie de Fouille de Textes (FdT), une discipline scientifique au croisement des statistiques, de l'analyse de données, de l'apprentissage, de l'Intelligence Artificielle (IA) et du Traitement Automatique de la Langue (TALN)), ont reçu une attention particulière. Cet intérêt est motivé d'une part par les besoins urgents de la part des biologistes et d'autre part grâce au succès rencontré dans les applications de FdT à Internet (par exemple, les moteurs de recherche en-ligne).

Dans cette thèse, nous présentons un système complet d'acquisition de connaissances à partir de textes, à base de techniques de FdT, et appliqué à une problématique particulière de la biologie moléculaire : l'étude de la régulation de l'expression des gènes chez l'homme. Nous proposons un ensemble de méthodes afin d'extraire, d'identifier et de structurer les données des publications de biologie moléculaire relatives aux réseaux de régulation de la transcription des gènes. Nous nous positionnons dans cette étude à deux échelles distinctes : d'une part à l'échelle du noyau de la cellule biologique et d'autre part, d'un point de vue intégratif, à l'échelle de l'organisme. Nous concevons ce travail comme un effort de confrontation d'une problématique importante de biologie moléculaire avec des méthodes et des solutions informatiques.

La première partie de ce mémoire propose de décrire le contexte du travail, à la fois dans sa composante biologique et dans sa composante informatique.

Le premier chapitre est consacré à la description succincte du processus biologique de la transcription de gènes et de son fonctionnement au sein de réseaux de régulation. Le but de cette section est de fournir une lecture générale des concepts biologiques qui seront manipulés dans le reste du document. Dans un premier temps nous allons définir les notions de fonctions biologiques et préciser notamment comment celles-ci sont régulées au sein de réseaux intégrés. Dans un deuxième temps nous proposerons un aperçu des mécanismes moléculaires impliqués dans l'expression des gènes ainsi que les systèmes de régulation mises en place par la cellule eucaryote. Finalement, nous illustrerons notre propos grâce à l'étude des modalités de la régulation de la transcription d'une famille de gènes exemple : les cytokines.

La deuxième partie propose un survol de la discipline de FdT. Nous mettrons en lumière quelques problématiques classiques de la FdT et les processus mis en œuvre afin d'y répondre. Nous centrerons alors notre propos sur la place de la FdT dans les questions de

biologie. Nous présenterons les différentes applications communément développées dans le contexte de la biologie ainsi qu'un tableau général de l'état de l'art du domaine. Cette partie mettra en lumière deux axes majeurs de la FdT appliquée à la biologie que sont d'une part la Reconnaissance des Entités Nommées (REN) et l'Extraction d'Information (EI). Nous préciserons alors tout au long de notre discussion les difficultés spécifiques aux textes de biologie que nous opposerons à celles rencontrées dans les approches "traditionnelles" de la FdT.

La deuxième partie du document concerne les méthodes et les résultats obtenus durant cette thèse. Dans une courte introduction nous résumerons la problématique du travail effectué. Nous présenterons notamment un modèle de la régulation de gènes qui nous permettra de définir le cadre de notre approche de FdT dans le contexte de l'information à extraire des textes.

Le troisième chapitre et le quatrième chapitre de ce manuscrit mettront en perspective les différentes méthodes développées avec les résultats des performances mesurées. Le troisième chapitre sera consacré spécifiquement aux solutions apportées afin de détecter et d'identifier les objets biologiques manipulés dans les textes (REN). Le quatrième chapitre proposera lui de répondre à la question de la découverte des relations d'intérêt qui animent les différents objets biologiques isolés des textes (EI). Tout au long de cette partie nous détaillerons les différentes phases de notre processus global de FdT dans l'ordre séquentiel de leur utilisation afin de répondre à notre problématique.

Nous conclurons alors ce document en dressant le bilan du travail effectué que nous mettrons en perspective.

Chapitre 1

La régulation de gènes

L'expression d'un gène est un processus biologique qui permet la production d'acides ribonucléiques (ARN) puis de protéines grâce l'information contenue dans un gène, telle que la séquence d'acide désoxyribonucléique (ADN).

La régulation de l'expression des gènes (ou régulation des gènes) se réfère au contrôle, par la cellule vivante, de la quantité et du rythme des changements de forme du produit fonctionnel d'un gène. Différentes étapes dans l'expression des gènes peuvent être modulées comme nous le verrons un peu plus loin dans ce document. Cette régulation donne à la cellule la possibilité de contrôler sa structure et sa fonction. Elle est à la base de mécanismes biologiques tels que la différentiation cellulaire et la morphogénèse et permet aux organismes de s'adapter et d'être polyvalents. Les protéines, remplissant une fonction particulière, ne sont produites que lorsque nécessaire.

Parmi tous ces points de contrôle mis à la disposition de la cellule, les mécanismes dépendant directement du processus de transcription et notamment l'initiation de la transcription sont les modes de régulation privilégiés chez les eucaryotes [CN07]. C'est cette forme particulière de régulation qui sera décrite en détail dans ce chapitre. Par exemple, nous passerons sous silence les mécanismes de contrôle traductionnels et post-traductionnels qui interviennent lors de la production des chaines polypeptidiques des protéines et de leurs conformations.

Il est aussi à noter que les modalités de la transcription et de sa régulation chez les organismes procaryotes et eucaryotes différent sur de nombreux points. Chez les procaryotes

il n'y a pas vraiment de noyau, alors que chez les eucaryotes il est très clairement définit. Ainsi, les procaryotes peuvent coupler transcription et traduction, mais les eucaryotes non. La transcription chez les eucaryotes se passe dans le noyau et la traduction se fera dans le cytoplasme. Chez les procaryotes un seul enzyme effectue la transcription pour tous les types d'ARN, tandis que chez les eucaryotes, trois enzymes distincts interviennent. Malgré l'intérêt certain des modèles de transcription procaryotes dans l'élucidation de certains mécanismes communs de régulation, nous nous intéresserons plus particulièrement à la transcription des gènes codant les chaînes polypeptidiques chez les eucaryotes dans cette section.

Dans une première partie introductive de ce chapitre nous montrerons le rôle des protéines dans les fonctions de la cellule et de l'organisme et notamment comment ces fonctions sont intégrées au sein de réseaux. Dans une deuxième partie, nous centrerons notre propos sur des réseaux spécifiques que sont les réseaux de la régulation de gènes. Nous exposerons alors avec détails le mécanisme de la transcription de gènes et les procédés par lesquels elle est régulée. Nous terminerons le chapitre en illustrant la complexité de la régulation de gènes en prenant l'exemple d'une famille de gènes particulière : les cytokines.

1.1 Fonctions et réseaux de régulation

Les protéines forment les briques élémentaires du vivant et participent à des processus biologiques très différents. A l'échelle moléculaire, les protéines remplissent les rôles essentiels à la vie. Chaque protéine est généralement spécialisée pour accomplir une tâche spécifique telle que, par exemple,

- la catalyse des réactions chimiques,
- l'acheminement des ions et des petites molécules à travers l'organisme,
- la communication intercellulaire et le fonctionnement coordonné de l'organisme,
- la défense de l'organisme contre les agressions externes,
- le mouvement ou la structuration de la cellule,
- etc.

Les protéines sont créées dans la cellule grâce à l'information contenue dans les séquences d'acides désoxyribonucléiques (ADN). L'information spécifique à une protéine est codée dans une portion particulière de l'ADN cellulaire nommée gène. L'ensemble des gènes d'un organisme est appelé génome. Chaque cellule d'un organisme (à quelques nuances près) comprend l'intégralité du génome et est stocké dans un compartiment particulier

qui est le noyau cellulaire. La synthèse d'une protéine à partir d'un gène est un processus hautement régulé. Dans la plupart des organismes, le gène est tout d'abord copié sous une forme intermédiaire dîtes ARNm (pour acide ribonucléique messager) lors d'une étape appelée transcription. Celui-ci va alors subir une phase de maturation, et être débarrassé de portions qui ne correspondent pas aux plans de constructions de la protéine telles que les introns, ou être éventuellement dégradé. L'ARNm mature peut ensuite être transporté hors du noyau chez les eucaryotes et subir la traduction. La traduction est l'étape finale de la synthèse de la protéine. Les ribosomes, avec l'aide de molécules tierces nommées ARNt (pour acide ribonucléique de tranfert), utilisent l'ARNm en tant que plan de construction de la protéine et procèdent à la mise bout à bout des briques de base de la structure protéique que sont les acides aminés. La protéine nouvellement synthétisée n'est généralement pas encore active et doit être encore modifiée afin de remplir sa fonction propre. Des exemples de modifications post-traductionnelles sont la mise en place de la structure tri-dimensionnelle et l'adjonction de groupements chimiques sucrés ou lipidiques.

Néanmoins, la compréhension des différentes étapes de la production des protéines ne permet pas d'élucider les différences de fonction constatées entre les cellules et le caractère conditionnel des rôles joués par les protéines synthétisées. Les gènes et leurs produits fonctionnent différemment dans les différentes cellules d'un organisme. Seul un certain nombre et non l'ensemble complet des protéines codées par le génome sont fabriquées par une cellule. De plus, il n'est pas rare qu'une même protéine, synthétisée par deux cellules différentes, ne remplisse pas le même rôle. Par exemple, la protéine BMP4 sert à transformer certaines cellules en os. Dans d'autres cellules, la protéine servira à modifier ces cellules en épiderme. Dans un troisième ensemble de cellules, BMP4 induira la division cellulaire alors que dans une quatrième lignée cellulaire, la protéine provoquera la mort cellulaire. L'analyse de la fonction d'un gène nécessite non seulement la connaissance de tous les constituants de la cellule mais encore de déterminer leurs interactions. Autre point capital, ces interactions au sein de la cellule varient non seulement en fonction de l'espace mais aussi du temps.

1.1.1 la fonction biologique

Une fonction biologique est une notion complexe. La définition d'une fonction dépend du niveau de détail dans lequel on se place (moléculaire, cellulaire, organisme) et de l'intérêt que l'on y porte. Karp [Kar00] donne une définition rigoureuse de la fonction

biologique et incorpore l'idée que celle-ci doit prendre en compte les différents niveaux de points de vue simultanément : d'une part, au niveau moléculaire, où la fonction biologique est perçue comme l'**action** que l'entité biologique exerce au sein de la cellule et d'autre part, au niveau cellulaire et supérieur, où elle est interprétée en tant que rôle joué par l'entité biologique au sein du fonctionnement de la cellule. En d'autres termes, la deuxième définition de la fonction biologique précise la contribution de l'entité biologique au comportement de la cellule. Ces deux notions sont très différentes et répondent à deux problématiques distinctes, l'auteur détaille ces deux aspects de la fonction biologique en les nommant *fonction locale* et *fonction intégrée* respectivement. L'auteur illustre cette différence en précisant qu'une même protéine peut posséder à la fois une fonction locale identique et une fonction intégrée différente dans deux organismes distincts. Soit par exemple, un enzyme E qui catalyse la même réaction dans deux bactéries différentes B1 et B2. Néanmoins, cet enzyme peut être utilisé dans la glycolyse dans B1 alors qu'il intervient dans la néoglucogenèse dans la bactérie B2. La fonction intégrée de E est donc variable tout en ayant une fonction locale identique entre B1 et B2. Ce contexte représente les multiples possibilités d'interactions entre gènes ou produits de gènes. Dans l'exemple précédent la différence observée du point de vue de la fonction intégrée de E peut être due, par exemple, au contexte de l'expression du gène de E dans les deux organismes. Ou bien encore parce que les deux mécanismes (glycolyse ou néoglucogénèse) n'interviennent pas simultanément au sein d'une même bactérie. Ces interactions, physiques ou non, sont le synonyme de la diversité au niveau de la cellule et cette diversité peut être étendue à un organisme tout entier où des millions de cellules interagissent.

1.1.2 les réseaux biologiques

La biologie des systèmes est une discipline de la biologie qui se focalise sur l'étude systématique des interactions complexes au sein des systèmes biologiques. L'organisme est perçu comme un réseau intégré et d'interactions entre les gènes, les produits de gènes et les autres composants cellulaires. Une définition étendue peut inclure dans cette liste les relations entre systèmes biologiques dont la physiologie est coordonnée ainsi que les facteurs environnementaux. La cartographie complète du génome humain [KSF+02] fournit un terrain de jeu idéal pour la biologie des systèmes car le catalogue relativement complet des composants génétiques de l'homme est disponible, en revanche les relations qu'ils établissent entre eux restent encore a élucider. De nombreux phénomènes biologiques tels que la formation des organes et plus tard l'établissement de leurs fonctions ne peuvent être

expliqués par l'action d'un gène unique ou de l'addition des actions de plusieurs gènes. Les acteurs de la transcription de l'ADN, de la maturation et la traduction de l'ARN sont identifiés mais la connaissance des interactions entre ces protéines et les séquences d'acide nucléique est limitée. De récents progrès expérimentaux ont été achevés dans ce domaine et la quantité d'information relative à cette problématique croit de manière soutenue depuis maintenant quelques années [MKMF$^+$06]. La biologie des systèmes demande aussi que les communautés de biologistes interagissent et mettent en commun leurs connaissances afin d'intégrer leurs données et suggérer de nouvelles expérimentations.

En génomique, deux types de réseaux se distinguent classiquement [T.07] : d'une part les réseaux physiques d'interaction et d'autre part les réseaux génétiques. Le premier définit l'architecture de la cellule en terme d'associations entre molécules (les plus étudiées sont les interactions protéine-protéine et protéine-ADN), de voies de signalisation et de transduction et d'autres machineries cellulaires. La notion d'interaction recouvre des réalités biologiques diverses :

- la régulation transcriptionnelle (par exemple la possibilité de fixation d'une protéine sur l'ADN),
- la régulation post-transcriptionnelle (par exemple, le contrôle de la dégradation de l'ARNm),
- la modification post-traductionnelle (par exemple la glycosylation des protéines),
- la formation de complexes protéiques,
- etc.

Le deuxième type d'interactions, en revanche, prévoit les relations fonctionnelles entre gènes et permet de mettre en relation le phénotype avec les mutations des gènes. L'intégration de ces deux types de réseaux à grande échelle est un pas nécessaire vers la description du fonctionnement de la cellule. La complémentarité et l'intégration de réseaux physiques et génétiques a notamment permis la compréhension de nombreuses voies de signalisation (voir [T.07] pour des exemples). A moyenne et grande échelle, l'intégration des réseaux ne peut être réalisée sans l'appui de méthodes automatisées. L'inférence de réseaux de régulation intracellulaires est rapidement devenu un des axes majeurs de recherche en bioinformatique [KCE04, WCBL02].

1.1.3 Caractéristiques des réseaux biologiques

Les réseaux d'interaction de différents types partagent un ensemble de principes qui établissent leurs relations.

Dynamicité Alors que les génomes peuvent être considérés comme étant globalement statiques, les réseaux d'interactions eux sont entièrement dynamiques et dépendent du contexte. Les interactions ne peuvent avoir avoir lieu que dans un type particulier de cellule, lors de certains stades du développement uniquement, etc.

Présence ou absence de direction dans l'interaction Une interaction peut être orientée dans le sens où il existe une direction biologique (cause et effet). Par exemple, une interaction entre un facteur de transcription et un promoteur est orientée, le facteur de transcription régulant le gène en amont et non l'inverse. En revanche une interaction protéine-protéine ne l'est pas, il n'y a pas de de relation de cause à effet biologique sous-jacente.

Les interactions stables et transitoires Des exemples d'interactions transitoires sont les réactions enzymatiques en général et la liaison d'un facteur de transcription sur un promoteur. L'interaction est ici provisoire et réversible. Sa durée est extrêmement variable selon le type d'interaction établie. Les interactions entres les protéines du cytosquelette et les pores nucléaires sont elles considérées comme étant stables. Une fois établies, elles sont permanentes.

Les interactions intégrées dans une voie de signalisation ou "de jonction" Selon Kelley *et al.* [KI05], une distinction peut être effectuée entre les interactions incluses dans une même voie de signalisation et celles qui permettent de connecter deux voies de signalisation apparentées. Si l'on se place du point de vue des interactions génétiques, les interactions "de jonction" peuvent relier des gènes possédant des fonctions redondantes ou complémentaires à d'autres. La délétion d'un de ces gènes provoque la perturbation d'un des deux réseaux connectés mais non les deux, à la différence des interactions intégrées dans une voie de signalisation. L'analogie est similaire pour les interactions de type physique.

Les interactions de première et de seconde catégorie Ici, si l'on considère deux molécules A et B dans un contexte d'interactions physiques, lorsque A et B interagissent par le biais de molécules intermédiaires, une interaction secondaire est ainsi définie. Il est possible aussi, qu'en sus A et B interagissent directement et soient liées par une interaction de type primaire. Les deux notions ne sont pas indépendantes et les interactions de seconde catégorie sont constituées d'interactions de première catégorie. Les interactions de deuxième catégorie permettent, dans un cadre expérimental, de prédire l'existence de nouveaux composants et d'interactions au sein de complexes protéiques lorsque les preuves d'interactions primaires n'ont pu être démontrées dans un premier temps [YPTM06].

1.2 Les réseaux de régulation de la transcription

Dans une certaine mesure, les cellules biologiques peuvent être assimilées à des "sacs" contenant un mélange hétérogène de composés chimiques. Lorsque l'on se positionne dans une contexte de régulation de gènes, ces composés chimiques sont les ARNs et les protéines qui résultent de l'expression des gènes. Les ARNs et les protéines interagissent alors ensemble avec différents degrés de spécificité. Certains diffusent dans la cellule. D'autres se lient à la membrane cellulaire afin d'interagir avec les molécules de l'environnement. D'autres encore traversent la membrane cellulaire et médient des signaux à longue portée dans les organismes pluricellulaires. L'ensemble de ces molécules et de leurs interactions composent alors un réseau de régulation de gènes.

Les quantités de la plupart des produits de gènes sont modulées via la contrôle de la transcription. D'autres points de contrôle de la production des protéines et de leur fonction sont notamment, et comme évoqué précédemment, la modification de la conformation des protéines ou encore l'adjonction de groupements chimiques au squelette protéique (suite à une phosphorylation transitoire ou de manière permanente grâce à l'addition d'un groupement glycosylé). Un des mécanismes majeurs responsables de cette régulation fine au niveau de la transcription des gènes résulte de l'interaction de facteurs protéiques spécifiques appelés facteurs de transcription sur l'ADN. Néanmoins de nombreuses zones d'ombre restent en suspens et notamment comment l'expression d'un génome entier peut être régulé par un nombre considérablement restreint de facteurs de transcription.

Dans de tels réseaux de régulation, les gènes peuvent être visualisés en tant que 'nœuds' de graphes dirigés dont les entrées sont des protéines telles que les facteurs de transcrip-

tion et les sorties le niveau de l'expression. Le nœud en lui-même représente une fonction qui peut être obtenue en combinant les fonctions d'entrées. Dans un réseau de type Booléen [GL07], les paramètres sont réduits à des fonctions binaires et les interactions sont modélisées comme des fonctions logiques.

Dans cette section, nous décrirons le processus de la transcription chez les eucaryotes et le rôle des facteurs de transcription. Nous élargirons ensuite le propos en introduisant quelques notions sur la structuration de ces réseaux.

1.2.1 La transcription

Les gènes existent dans le noyau cellulaire sous la forme d'une structure complexe tri-dimensionnelle, appelée chromatine. Elle est composée de la répétition de nucléosomes qui contiennent chacun un cœur de protéines histones (H2A, H2B, H3 et H4, organisées sous la forme d'un octamère globulaire 2x(H2A, H2B, H3, H4)) autour duquel 145 paires de bases d'une hélice double brin d'ADN s'enroulent deux fois. cette structure dite en "bobine" [KL99] est compactée sous la forme de fibres de chromatine, elles mêmes structurées pour former les chromosomes. La chromatine condensée est généralement vue comme une architecture hermétique à la transcription des gènes qu'elle contient. En revanche, lorsque la chromatine est décondensée ou "remodelée", la transcription de gènes devient possible (les mécanismes sont décrits dans [AD00]). En effet, l'ADN embarqué dans le noyau de la cellule est formidablement empaqueté "par défaut", c'est-à-dire dans son état initial ou de base. Pourtant le patrimoine génétique peut être accédé spécifiquement dans chaque cellule. Il est suggéré, qu'au minimum, une dizaine de milliers de gènes soit présent dans le génome de la plupart des vertébrés. Or, dans les cellules la grande majorité des gènes n'est pas exprimée. L'état "par défaut" de la chromatine est ainsi réprimé et seule la perturbation locale de facteurs répressifs peuvent permettre l'activation de gènes dans un contexte spécifique à une cellule [Wei85]. Les changements dans la structure de la chromatine peuvent survenir en réponse à des signaux relatifs au développement cellulaire et sont alors maintenus dans la cellule différenciée. Des modifications de la chromatine peuvent aussi avoir lieu suite à la présence de *stimuli* spécifiques et dans ce cas les changements effectués sur la chromatine sont réversibles : elle retrouve son état initial dès que les signaux ayant provoqués sa transition ont disparu. Deux mécanismes critiques semblent permettre le remodelage de la chromatine. D'une part, des complexes de remodelage de la chromatine ATP-dépendants (des protéines apparentées à SWI/SNF) ont été caractérisés

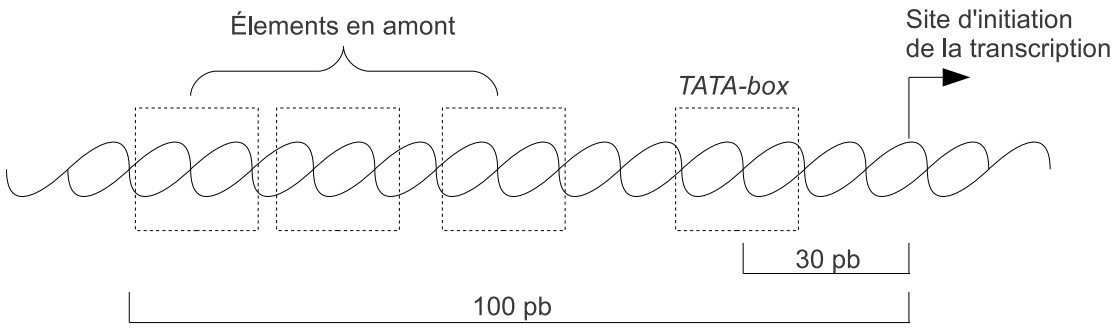
chez *E. Coli*, *S. Cerevisiae*, *C. Elegans* et de nombreux mammifères [A.99]. Ce sont des complexes multi-protéiques qui utilisent l'énergie libérée lors de l'hydrolyse de l'ATP pour rompre la structure compactée de la chromatine et ainsi permettre l'accès aux complexes transcriptionnels à l'ADN. D'autre part, il existe des co-activateurs de la transcription qui possèdent une activité histone acetyl tranférase (HAT) et des co-represseurs de la transcription avec une activité histone déacétylase (HDAC) [PCP00]. Les histones possèdent une extrémité N-terminale située à l'extérieur du cœur du nucléosome avec des lysines sensibles à l'actéylation [KL99]. Lorsque les histones sont déacétylés, les nucléosomes sont compactés et la transcription est difficilement réalisable. Une fois acétylés, la compaction est plus lâche et les facteurs de la transcription peuvent plus facilement atteindre l'ADN. Les régions hyperacétylées sont associées aux gènes les plus actifs alors que l'hypoacétylation des histones est retrouvée dans les zones réprimées de la chromatine. D'autres modifications des extrémités N-terminales des histones telles que la phosphorylation et la méthylation peuvent jouer un rôle restreint dans le mécanisme de l'accessibilité à la transcription [PKW$^+$00]. L'ensemble des modifications des queues d'histone agissent de concert et il est proposé que certains facteurs de transcription et d'autres protéines peuvent s'y lier [PKW$^+$00].

Les complexes de facteurs de transcription s'assemblent sur les zones promotrices des gènes et servent souvent à recruter des co-activateurs, des composants de la machinerie transcriptionnelle à l'état basal ainsi que le complexe enzymatique Polymerase II [Car98]. Ces grands complexes multi-protéiques forment l'*enhanceosome*. Certains facteurs de transcription sont capables, semble-t-il, de reconnaitre leurs sites de liaison sur la chromatine et de recruter la machinerie de modification de la chromatine de manière gène-dépendante [MTK99]. D'autres facteurs de transcription peuvent alors interagir avec l'ADN une fois que la chromatine a été remodelée.

1.2.1.1 Structure des promoteurs de gènes

Les promoteurs de gènes qui permettent la transcription de larges quantités d'ARNm ont des structures similaires [S.97]. Ils possèdent une séquence TATA (aussi connue sous le nom de *TATA-box* ou séquence de Goldberg-Hogness) située environ 30 paires de bases en amont du site où la transcription démarre. En aval de la séquence TATA existe un ou plusieurs autres éléments du promoteur (Figure 1.1). Uniquement certaines paires de bases sont nécessaires à la transcription du gène. Des gènes clonés peuvent être transcrits lorsqu'ils sont placés dans des noyaux d'ovocytes de grenouille ou des fibroblastes

Fig. 1.1 – Région promotrice de gène d'eucaryote typique



La séquence promotrice contient une *TATA-box* et différents autres élements en amont.

ou lorsqu'ils sont incubés en présence d'ARN Polymérase II et de nucléotides et d'extraits nucléaires. Lorsque la transcription effective d'un gène est confirmée, des enzymes de restriction permettent de découper spécifiquement des zones du gène ou des portions alentours. Le but de la manipulation étant de détecter si le gène peut être encore transcrit avec efficacité. Selon les perturbations observées sur la transcription ou le rétablissement de l'expression et à force de raffiner ces zones de découpage, il est possible d'isoler et d'identifier des éléments particuliers dans la zone promotrice. Ces éléments peuvent faciliter, entraver, voire être nécessaires ou encore supprimer totalement la transcription d'un gène. La topographie d'un promoteur type chez un eucaryote peut être succinctement résumé en deux zones, d'une part le promoteur proximal (directement en aval de la séquence TATA) et d'autre part le promoteur distal (où sont situées les séquences dites *enhancers* ou activatrices et les séquences *silencers* ou répresseurs).

Fonction du promoteur Les promoteurs doivent intéragir avec l'ARN Polymerase II pour produire la transcription d'un gène. Il existe trois ARN Polymérases chez les organismes eucaryotes, possédant des structures et des propriétés différentes. L'ARN Polymérase II est responsable de la production d'ARNm (ou ARN messagers) précurseurs (ou non- matures). C'est l'ARNm qui servira de plan pour la construction des protéines de la cellule. Il est intéressant de noter que, chez les eucaryotes, aucune ARN Polymérase ne peut se lier seule avec efficacité sur l'ADN. La fixation préalable sur l'ADN de facteurs de transcription et leur interaction consécutive avec l'ARN polymérase sont nécessaires. Certains facteurs de transcription permettent à l'ARN Polymérase de commencer la transcription au bon endroit sur l'ADN et de ne le faire qu'à certains moments et uniquement

dans certaines cellules. De plus, un rôle a été proposé pour la disponibilité des gènes grâce aux changements de conformation de la chromatine dans la spécificité cellulaire de la transcription. Les profils d'expression dépendants du type cellulaire sont établis lors de la différenciation cellulaire. Par exemple, le gène de l'IL2 peut subir des modifications lors du développement qui lui permet d'être exprimé dans les cellules T matures et uniquement dans le thymus [MGE01].

Mécanisme de la transcription Il a été montré qu'un minimum de six protéines nucléaires est nécessaire afin d'initier la transcription par l'ARN Polymérase II :

- Tout d'abord le complexe TFIID se fixe à la séquence TATA. Cette étape sert de point de départ à la construction du complexe transcriptionnel et permet d'empêcher la stabilisation des nucléosomes dans la région promotrice en bloquant leur production. La fixation de TFIID sur la séquence TATA est facilitée et stabilisée par le facteur de transcription TFIIA.
- TFIID fixe alors TFIIB qui se lie à son tour directement à l'ARN Polymérase II. Il est à noter que la Polymérase se lie aussi, et consécutivement, à TFIID par le biais de son extrémité carboxy-terminale. de nombreux facteurs de transcription vont alors pouvoir soit faciliter, soit empêcher la transcription d'un gène spécifique en modifiant la stabilisation de TFIID et TFIIB sur le complexe déjà formé.
- Soit avant ou pendant sa fixation avec TFIIB, l'ARN Polymérase II s'associe à TFIIF et TFIIE. TFIIF possède une activité enzymatique capable de "dérouler" l'hélice d'ADN et ainsi rendre possible la transcription. Le rôle de TFIIE semble quant à lui d'être la production de l'énergie nécessaire à la transcription.
- A ce moment, TFIIH va permettre la libération de l'ARN polymérase II au complexe formé et ainsi de rendre à l'enzyme sa mobilité en phosphorylant la queue de l'ARN Polymérase II liée à TFIID.
- TFIID est une protéine multi-mérique et seule une sous-unité (nommée TBP, pour *TATA Binding Protein*) se lie à la séquence TATA. Quelques unes des autres sous-unités (appelées TAFs, pour *TATA-binding protein-associated factors*) permettent, d'une part de déterminer si TFIID reste sur la séquence TATA, et d'autre part d'agir en tant que co-activateurs et ainsi relier les facteurs de transcription facilitateurs de la transcription au complexe. Voir la figure 1.2 pour illustration.
- Il existe certains promoteurs sans séquence TATA mais qui néanmoins utilisent l'ARN Polymérase II. Dans ces cas, d'autres protéines interagissent avec la zone promotrice. Ce sont généralement des protéines qui vont alors directement lier TFIID,

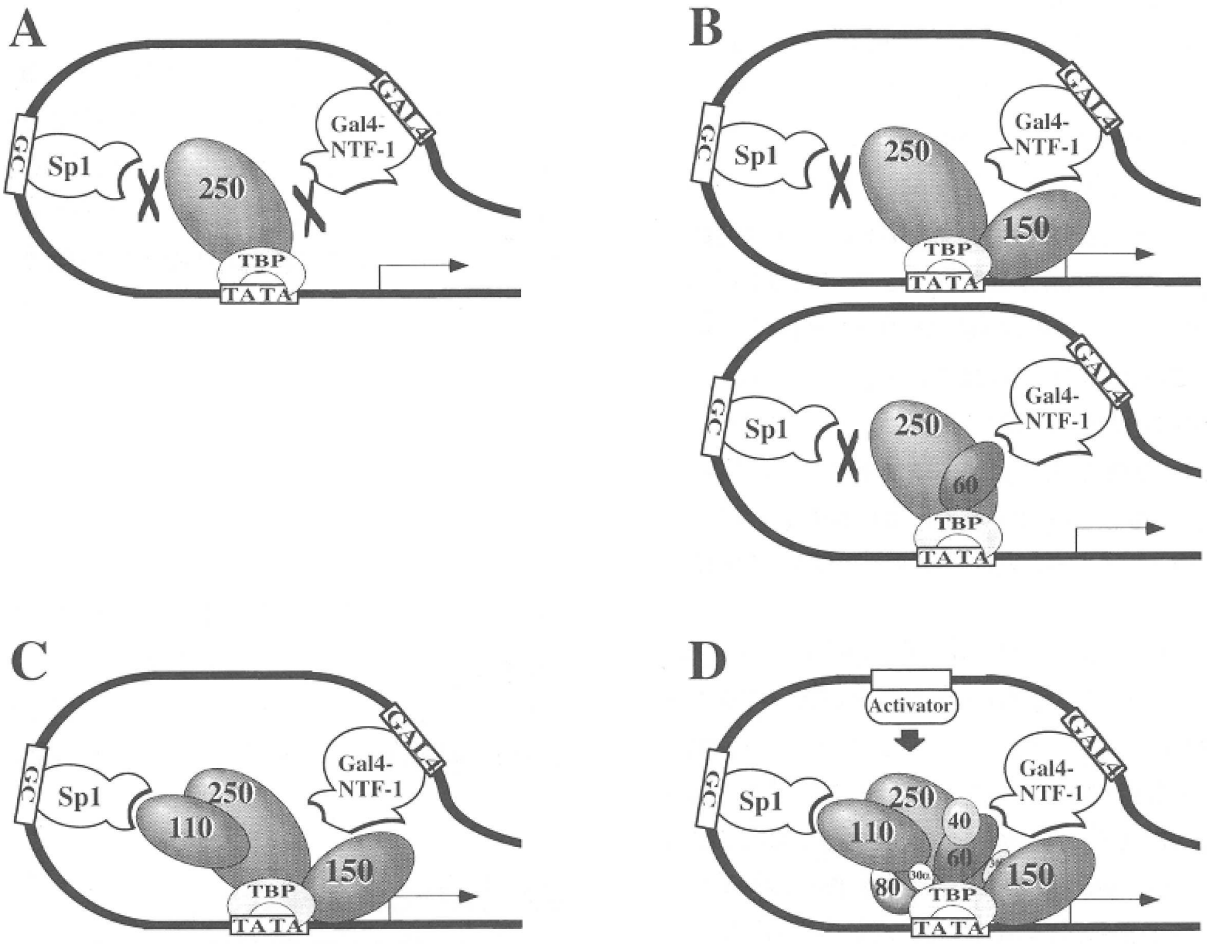
ou par l'intermédiaire d'une TAF, et la cascade de construction du complexe transcriptionnel peut avoir lieu normalement. Par exemple, SP1 peut interagir avec TFIID sur un promoteur sans séquence TATA. Cette protéine se lie sur l'ADN via des séquences riches en guanines et en cytosines.

Lorsque la TBP n'est plus liée au promoteur, la transcription du gène ne peut avoir lieu. Deux TAFs (de 250 et 150kDa) sont cruciaux dans la détermination de cette liaison [JLC+94]. Ces TAFs reconnaissent spécifiquement des éléments en amont du promoteur, qui, lorsqu'ils sont présents, permettent soit de stabiliser soit de déstabiliser la TBP sur le promoteur. L'association de la TBP avec différents TAFs permet au complexe transcriptionnel d'être activé par des protéines liées à des séquences activatrices ou en amont du promoteur. De plus, différents TAFs sont capables d'être co-activés grâce à des facteurs activateurs de transcription non-TAFs. Ce sont ces facteurs particuliers auxquels nous réservons par la suite le nom de facteurs de la transcription. Par exemple, et comme illustré précédemment, SP1 fait partie de la classe des facteurs de transcription et est très souvent rencontré sur les promoteurs eucaryotes [KCMT87].

1.2.1.2 Les facteurs de transcription

Les facteurs de transcription sont des des protéines qui peuvent interagir sur les régions activatrices ou d'autres zones particulières de l'ADN d'un promoteur. Leurs interactions sont telles que pour n'importe quelle cellule de l'organisme la transcription n'est uniquement possible que dans un petit nombre de promoteurs. La plupart des facteurs de transcription peuvent se lier sur des séquences particulières de l'ADN et sont généralement classés en familles de similarités de structure [Win97]. A l'intérieur de chaque famille de facteurs de transcription, les protéines partagent une structure commune de sites de liaison à l'ADN. Néanmoins des différences subtiles dans la composition en acides aminés de ces sites d'interactions entre chaque membre d'une même famille font que les séquences d'ADN reconnues sont parfois différentes. En plus de ces sites de liaison à l'ADN, les facteurs de transcription possèdent un domaine d'activation de la transcription. Ce domaine permet au facteur de transcription d'interagir avec d'autres protéines impliquées dans la liaison de l'ARN Polymérase II au promoteur. Cette interaction a souvent pour but d'augmenter l'efficacité de la construction du complexe transcriptionnel et de sa liaison à l'ARN Polymérase II. Quelques facteurs de transcription représentatifs ainsi que les séquences nucléotidiques sur lesquelles ils interagissent sont illustrés dans le tableau 1.1.

Fig. 1.2 – Un modèle de l'activation de la transcription par différents facteurs de la transcription spécifiques au promoteur (tiré de Chen *et al.* [JLC+94])



[A], un complexe TBP/TAF minimal n'est pas suffisant pour initier la transcription. [B], deux complexes tri-mériques distincts peuvent initier la transcription grâce à Gal4-NTF-1 et non par SP1. [C], un complexe trimérique avec TFII110 peut médier une activation de la transcription grâce à SP1. [D], un complexe de protéine TFIID complet peut initier la transcription par le biais de très nombreux facteurs. La protéine activatrice en [D] a été stabilisée sur l'ADN par d'autres interactions.

Tab. 1.1 – Quelques facteurs de transcription représentatifs

Facteur	Motif nucléique consensus d'interaction	Commentaire
Myc et Max	$CACGTG$	Myc a été originellement identifié en tant qu'oncogène rétroviral. Max s'associe spécifiquement avec Myc dans les cellules.
Fos et Jun	$TGA\frac{C}{G}T\frac{C}{A}A$	Tous les deux identifiés originellement en tant qu'oncogènes retroviraux. S'associent ensemble dans les cellules. Aussi connus en tant que hétérodimère sous le nom d'AP1.
CREB	$TGACG\frac{C}{T}\frac{C}{A}\frac{G}{A}$	Se lie à l'élément CRE (cAMP Response Element). Famille d'au minimum 10 facteurs, provenant d'un épissage alternatif ou de différents gènes. Peut former des dimères avec Jun
ERBA ou TR (Thyroid Hormone Receptor)	$GTGTCAAAGGTCA$	Identifié originellement en tant qu'oncogène rétroviral. Membre de la superfamille des récepteurs aux hormones thyroïdes/stéroïdes. Se lie à l'hormone thyroïde.
ETS	$\frac{G}{C}\frac{A}{C}GGA\frac{A}{T}G\frac{T}{C}$	Identifié originellement en tant qu'oncogène rétroviral. Prédomine dans les cellules T et B.
GATA	$\frac{T}{A}GATA$	Famille de 6 facteurs spécifiques aux cellules érythroïdes.
MYB	$\frac{T}{C}AAC\frac{G}{T}G$	Identifié originellement en tant qu'oncogène rétroviral. Facteur spécifique des cellules hématopoïétiques.
MyoD	$CAACTGAC$	Contrôle la différenciation des cellules musculaires.
NF ?B et Rel	$GGGA\frac{A}{C}TN^1\frac{T}{C}CC$	Identifiés originellement de manière indépendante. Rel tout d'abord identifié en tant q'oncogène rétroviral. Prédominants dans les cellules T et B.
RAR (Retinoic Acid Receptor)	$ACGTCATGACCT$	Se lie aux éléments RAREs (Retinoic Acid Response Elements). Se lie aussi au site Jun/Fos.
SRF (Serum Response Factor)	$GGATGTCCATATTAGGACATCT$	La séquence nucléotidiques est présente dans beaucoup de gènes inductibles par les facteurs de croissance du sérum.

[1] N signifie que n'importe quelle base peut occuper cette position.
Cette liste propose uniquement une infime fraction représentative des centaines de facteurs identifiés.

Les familles de facteurs de transcription Un exemple[1] de classification basée sur la présence de domaines protéiques fonctionnels est mis à jour par la société **BIOBASE GmbH** qui administre la base de données de facteurs de transcription **TRANSFAC**[2]. Il existe de très nombreuses familles de facteurs de transcription, et ceux brièvement survolés ci-dessous en sont les principales.

- Les protéines à homéo-domaine. Ces protéines sont très importantes dans le règne animal pour la définition de l'axe antéro-postérieur lors du développement. L'homéo-domaine consiste en 60 acides aminés sous la forme du motif hélice-tour-hélice de sorte que la dernière hélice se retrouve dans le sillon majeur de l'ADN qu'il reconnait. Quelques exemples de protéines à homéo-domaine chez *D. Melanogaster* : Abdominal B, Bicoid, Engrailed, Even-skipped, Paired, Ultrabithorax.
- Les protéines POU. Certains facteurs de transcription possèdent à la fois un homéo-domaine et un deuxième site de liaison à l'ADN qui peut être le domaine POU [WRR+88]. Un exemple de protéine à domain POU est le facteur de transcription Pit-1. Pit-1 est retrouvé dans l'hypophyse. L'expression tissu-spécifique de l'hormone de croissance dans l'hypophyse antérieure est médiée par Pit-1. Lorsque les gènes d'hormones de croissance sont clonés et placés dans des extraits nucléaires de cellules non-pituitaires, ils ne sont pas transcrits. En revanche, les hormones de croissance sont exprimées si les extraits nucléaires proviennent de l'hypophyse antérieure ou si Pit-1 est injecté dans l'extrait [MM87].
- Les protéines à domaine bHLH. Les facteurs de transcription MyoD et Myogenin spécifiques aux muscles contiennent le domaine bHLH (pour *basic helix-loop-helix*) ainsi que de nombreuses protéines qui déterminent le développement des cellules du système nerveux périphérique de la drosophile telles que les protéines de Daughterless et de Achaete-scute. La détermination du sexe de la drosophile semble aussi liée à l'action de facteurs de transcription possédant ce motif. Les protéines bHLH agissent sous forme d'hétéro-dimères.
- Les protéines à domaine bZip. La structure des facteurs de transcription à domaine bZip (pour *basic leucine zipper*) est semblable à celle des protéines bHLH. Ce sont des dimères dont chacune des sous-unités contient à leur extrémité C-terminale un domaine basique de liaison à l'ADN suivi de plusieurs leucines au sein d'une hélice. Les leucines du motif sont positionnées dans l'hélice de telle manière qu'elle puissent interagir avec les leucines d'autres protéines bZip pour former une "fermeture à

[1] http ://www.gene-regulation.com/pub/databases/transfac/cl.html
[2] http ://www.biobase-international.com/pages/index.php ?id=transfac

glissière" et ainsi permettre la dimérisation. Proche du domaine bZip est retrouvé un domaine de régulation qui peut avoir un effet activateur ou inhibiteur sur l'expression de gènes. Des exemples de facteurs de transcription contenant le motif bZip sont AP1 et C/EBP.

- Les protéines à doigts de Zinc. Chaque protéine à doigts de Zinc possède de deux à plusieurs de ces motifs qui se placent dans les sillons majeurs de l'ADN lors de l'interaction. Ces domaines sont liés en tandem et sont stabilisés par un ion Zinc au cœur de chaque domaine grâce à deux cystéines à la base des hélices et deux histidines internes. Des exemples de protéines à doigts de Zinc sont Sp1, Egr-1 et Krüppel (chez la drosophile).
- Les protéines de courbure de l'ADN. Ces protéines ne possèdent pas de motifs fixes mais majoritairement contiennent le domaine HMG d'environ 80 acides aminés. Le rôle de ces protéines n'est pas de lier l'ADN contrairement aux autres facteurs de transcription évoqués plus haut mais de courber l'ADN afin de faciliter l'accès à d'autres activateurs ou répresseurs de la transcription. Ce sont ces protéines qui sont responsables de la formation de l'*enhanceosome*. Des exemples de protéines courbant l'ADN sont LEF-1 et SRY.

1.2.2 Les motifs des réseaux de régulation

La décomposition en modules fonctionnels permet de décrypter la structure logique de réseaux complexes. Les motifs des réseaux sont les briques élémentaires de construction des réseaux de régulation. Ils peuvent être comparés à des circuits récurrents d'interaction dont l'ensemble des inter-connections forme un réseau de régulation [MSOI+02]. Chaque motif transporte un type spécifique de fonction de traitement de l'information du réseau [Alo07]. Les descriptions des motifs les plus étudiés présentées ci-dessous ne prennent pas en considération les données cinétiques, pourtant essentielles à l'établissement de tels réseaux, et ceci dans le but de simplifier le propos.

Il existe deux types de réseaux qui répondent à deux besoins distincts. D'une part les réseaux de régulation de type "senseur", qui produisent rapidement une réponse à un *stimulus* et sont réversibles, et d'autre part les réseaux de régulation de type "développemental", qui à l'inverse répondent aux signaux à l'échelle du temps de vie de la cellule et orientent le développement de la cellule en produisant des effets irréversibles. La première catégorie de réseaux n'utilise exclusivement que les trois premiers motifs décrits alors que

la deuxième catégorie de réseaux est construite grâce à l'ensemble des motifs présentés. Voir la figure 1.3 pour illustration des différents motifs.

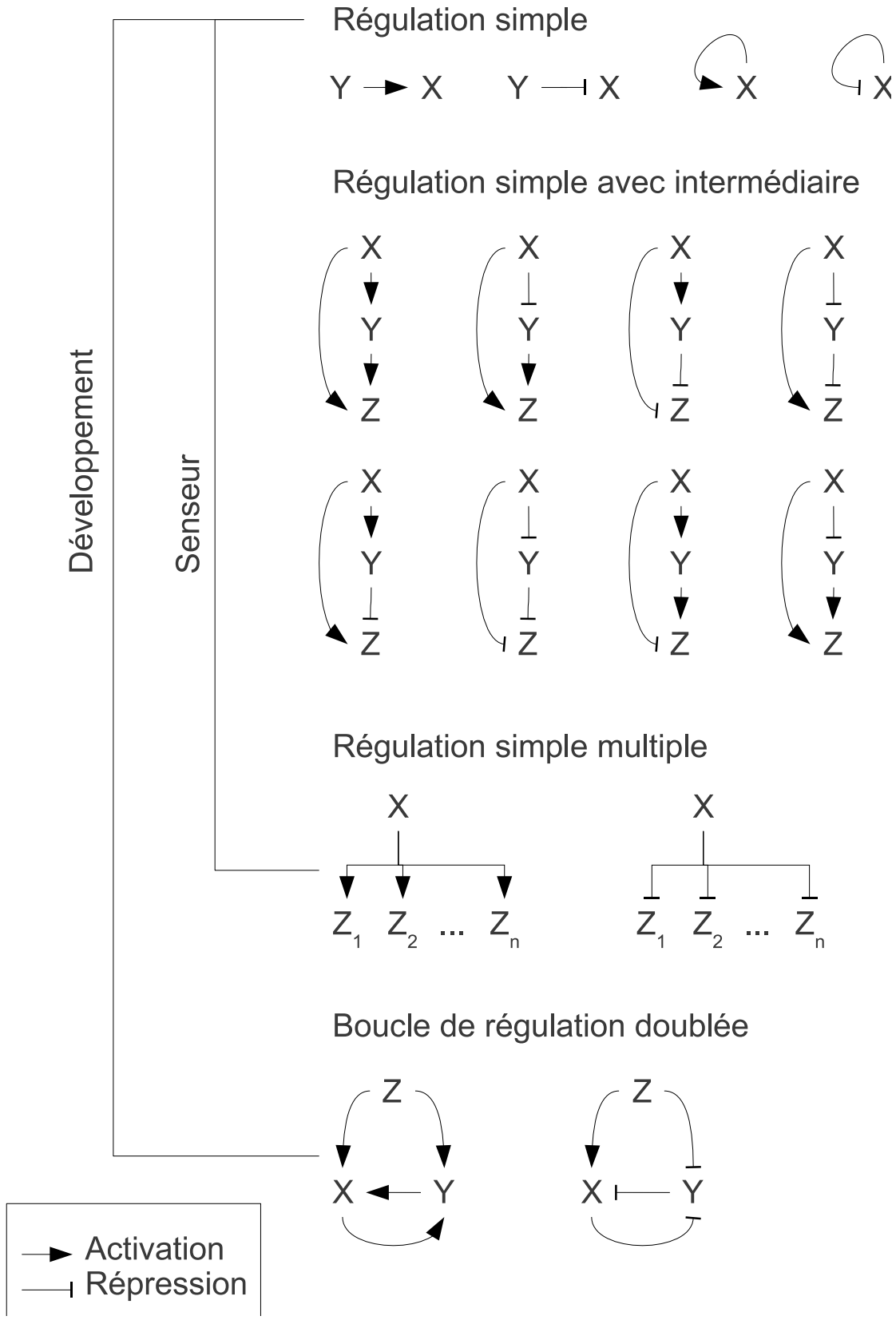
Régulation simple Le type d'interaction d'un réseau de régulation de la transcription le plus simple est la régulation directe d'un gène X par une facteur de transcription Y. En général, Y est activé par un signal, ce signal peut être chimique, mécanique, lumineux, etc. Lorsque la transcription commence, la concentration du gène X augmente et converge vers un plateau stable. Ce niveau stationnaire équivaut au rapport entre la production et la dégradation du produit de gène qui opèrent simultanément. La dégradation peut être un phénomène actif (via l'intervention d'enzymes) ou simplement du à l'effet de dilution résultant de la croissance cellulaire. Le phénomène d'auto-régulation négative est la capacité d'un facteur de transcription a réprimer son propre gène. C'est en général un mécanisme qui se met en place rapidement et intervient lorsque la concentration du produit de gène dépasse un certain seuil. L'auto-régulation positive est, à l'inverse, l'augmentation de la production du produit d'un gène codant un facteur de transcription par sa propre protéine. A la différence de l'auto-régulation négative, c'est un mécanisme lent.

Régulation simple multiple Ici un gène X possède une action sur plusieurs gènes Z sans intermédiaire. L'expression des gènes Z est ainsi coordonnée grâce à X. Les gènes Z possèdent alors des fonctions communes. Il est intéressant de noter que la protéine régulatrice X régule les gènes Z à des seuils différents et provoque une séquence temporelle d'activation ou de répression des gènes Z lorsque X est actif. Cette propriété résulte notamment des différences subtiles dans la séquence d'interaction de chaque promoteur Z pour la protéine X et du contexte (environnement) de cette interaction.

Régulation simple avec intermédiaire Ce type de motif a été bien étudié chez *E. Coli* et *S. Cerevisiae* et existe aussi chez l'homme. Ce motif est composé de l'interaction de trois gènes : un gène X qui régule un gène Y et un gène Z et Y qui régule Z. Les gènes X, Y et Z sont reliés par des interactions d'activation et de répression. Voir la figure 1.3 indiquant les combinaisons de régulation possibles. Les deux motifs décrits ici sont les plus retrouvés chez *E. Coli* et *S. Cerevisiae* :

- X et Y sont tous deux soient des activateurs de la transcriptions soient des répresseurs. X et Y peuvent agir simultanément et l'action de concert des deux gènes est alors nécessaire pour activer ou réprimer la transcription de Z. L'activité relative à

Fig. 1.3 – Motifs récurrents des réseaux de régulation de la transcription



Z possède alors un certain délai, du au fait que Y est nécessaire au même titre que X à la régulation de Z et que sa concentration doit être suffisante pour avoir un effet sur Z. En revanche l'arrêt de l'activité de Y ou X provoque un effet immédiat sur Z car les deux gènes sont nécessaires à sa régulation. Une autre possibilité est que seul X ou Y agit sur la transcription de Z à un moment donné. Le temps effectif de l'effet sur la régulation de Z est alors inversé par rapport à ce qui est exposé au dessus pour les raisons opposées.

– X et Y ont un effet opposé sur la régulation de Z et X régule différemment Y et Z. Si X active Z mais répresse Y alors Z est produit rapidement dès que X est actif. Toutefois, lorsque la concentration de Y dépasse un certain seuil et que le seuil de répression de Z est atteint, la concentration de Z décroit. On peut alors observer un effet de "pulse" sur la régulation de Z. De plus, si Y ne réprime pas complètement Z, la concentration de Z se stabilise à un niveau non nul. La vitesse de réponse est encore plus rapide que dans le mécanisme d'auto-régulation négative.

Il n'est pas rare que plusieurs de ces réseaux se combinent. X et Y ayant une action sur plusieurs gènes Z, avec très souvent des seuils d'activation ou de répression différents.

Boucle de régulation doublée Ici deux facteurs de transcription se régulent l'un l'autre. Soit les deux régulateurs s'activent mutuellement ou au contraire se répriment. Dans le cas des boucles de type activateur, les deux deux facteurs de transcription sont soit activés soit réprimés de concert. En revanche, dans le cas des boucles de type répresseur, seul un des deux régulateurs est activé. Une propriété intéressante de ce type de réseaux est l'effet de "mémoire" du signal rendu possible : même après l'arrêt du *stimulus* original, les deux facteurs de transcription continuent de se réguler mutuellement jusqu'à ce que les concentrations de leurs produits atteignent un plateau.

Cascade de transcription A la différence des réseaux de type "senseur", les réseaux liés au développement peuvent être particulièrement longs et posséder plusieurs couches de motifs mis en série [MSOI+02, Alo07]. Le temps de transmission du message au sein de ces cascades est assez lent et est consistant avec la fonction de détermination du destin de la cellule.

1.3 Famille de gènes d'intérêt : les cytokines

Les cytokines forment un groupe important de protéines qui fonctionnent an tant que messagers intercellulaires. Elles fonctionnent généralement dans un environnement proche du lieu de sécrétion en intéragissant avec des récepteurs spécifiques à la surface de la cellule. Les deux modalités d'actions, paracrine (la cytokine et son récepteur sont produits par la même cellule) et autocrine (la cytokine et son récepteur sont produits par deux types cellulaires différents), sont envisageables. Dans le système immunitaire, un large nombre de cytokines joue un rôle important à la fois dans le développement du système immunitaire et lors de la réponse de l'organisme à une infection. Les cytokines sont souvent groupées en familles en fonction de similarités de fonctions ou selon leurs structures protéiques [Mic05]. Des études *in vivo* et *in vitro* ont démontrées depuis les années 80 qu'individuellement ces protéines à la fois partagent souvent les mêmes rôles au sein du système immunitaire mais aussi possèdent des fonctions uniques.

1.3.1 La transcription des cytokines

De nombreuses cytokines sont spécifiquement exprimées par certains types cellulaires. Certaines ne sont exprimées que dans un type particulier de cellule (par exemple, l'IL2 est transcrite presque qu'exclusivement dans les cellules T matures) alors que d'autres montrent des spécificités très larges, mais non ubiquitaires. Par exemple, l'IL12 et IL18 peuvent devenir disponibles pour la transcription lorsque les macrophages et les cellules dendritiques maturent. Une deuxième caractéristique d'importance est que beaucoup de cytokines sont uniquement exprimées suite à l'exposition de la cellule à un ensemble particulier de signaux. Les régions régulatrices de ces gènes sont sensibles aux voies de transduction dépendantes des récepteurs membranaires de la cellule. Pour beaucoup de cytokines la spécificité cellulaire et la phase d'accessibilité des zones régulatrices du gène par les modifications de la chromatine doivent être coordonnés. Des régulations anormales de la transcription des gènes des cytokines dans le système immunitaire ont été observées dans de très nombreuses pathologies liées au système immunitaire et sont parfois l'élément clef du déclenchement de la maladie. Par exemple, des syndromes atopiques et divers évènements allergiques sont liés à l'expression non-physiologique de l'IL4, IL5 et de l'IL13 et les problèmes d'autoimmunité sont associés à la production de cytokines telles que l'IFNγ [AMS96]. La compréhension des mécanismes de la transcription des gènes des cytokines est ainsi une priorité dans le développement de traitements thérapeutiques dans

les maladies liées au système immunitaire.

Les zones promotrices et activatrices de nombreuses cytokines ont été largement étudiées depuis plus de dix ans (par exemple dans [JLR95, BMRJ+05, BH97]). Elles sont généralement composées de sous-régions denses en sites de liaison aux facteurs de transcription qui agissent de manière coopérative sur la transcription. Parmi les familles de facteurs de transcription les plus étudiés et se liant aux promoteurs de cytokines, nous retrouvons les protéines NFAT, NK-κB/Rel et BZip. Ces facteurs inductibles coopèrent avec des facteurs constitutifs ou des facteurs spécifiques à la cellule afin de constituer l'*enhanceosome* et d'activer la transcription.

1.3.1.1 Différentiation des cellules T CD4+

Les cellules T auxiliaires CD4+ se différencient en deux sous-ensembles de cellules immunitaires effectrices : Th1 et Th2. Ces deux lignées sont caractérisées par les types de cytokines qu'elles produisent et sécrètent. Les cellules Th1 sécrètent de l'IFN-γ, de l'IL2 et du TNF-β alors que les Th2 sécrètent de l'IL4, IL5, IL6, IL10 et IL13. Les deux lignées ont deux fonctions différentes au sein du système immunitaire de part cette spécificité de production de cytokines. D'une part les Th1 ont un rôle important dans les réponses immunitaires contre les agents infectieux intracellulaires (des dysfonctionnements peuvent alors provoquer des problèmes d'auto-immunité) et d'autre part les Th2 sont elles spécialisées dans la protection contre les agressions parasitaires (mais peuvent provoquer asthme et divers problèmes liés à l'allergie). Agarwal *et al.* [AR98] ont proposé un modèle de régulation de la transcription impliquant le remodélage de la chromatine afin d'expliquer cette différence de fonction entre les deux populations de cellules. Le modèle explique que l'accessibilité à la chromatine des *loci* des gènes des cytokines concernées est modifiée et permet ainsi leur activation conditionnelle. Ce changement de l'état de la chromatine est sous l'influence de différents *stimuli* comme le contact de la cellule avec certains antigènes, l'action de co-stimulateurs, le fond génétique de l'individu et surtout l'activité d'une signalisation dépendante de cytokines. L'influence, d'un côté, de l'IL4 et STAT4 et de l'autre, de l'IL-12 et STAT6 permettent de générer les populations Th1 et Th2, respectivement [RS99].

1.3.1.2 Expression inductible des cytokines

Lorsque les gènes de cytokines sont accessibles à la machinerie transcriptionnelle, de nombreux *stimuli* extra-cellulaires sont interprétés par les récepteurs cytoplasmiques qui à leur tour activent un nombre conséquent de voies de transduction cytoplasmiques et permettent l'expression des gènes de cytokines. La nécessité pour la cellule d'avoir recours à plusieurs voies de transduction pour l'activation d'un gène peut être perçue comme l'assurance d'une régulation fine mais permet aussi de limiter les risques que l'expression soit déclenchée suite à un dysfonctionnement quelconque. Deux mécanismes principaux sont à l'origine de l'activation de la transcription des gènes des cytokines. D'une part les combinaisons de fixation des facteurs de transcription sur les promoteurs et d'autre part le remodelage de la chromatine.

Modules En règle générale les sites de liaison aux facteurs de transcription sur les promoteurs de cytokines sont considérés comme faibles (les interactions établies entre la protéine et l'ADN sont trop faibles pour permettre la stabilisation du complexe) et un site unique ne peut généralement activer la transcription seul. De nombreux modules de deux ou plus sites de liaison ont été identifiés sur les promoteurs de cytokines. L'activation d'un promoteur est ainsi possible grâce à l'action coopérative de la fixation indépendante de facteurs de transcription sur ces sites. Par exemple, les promoteurs de l'IL2 [JMM+93], de l'IL3 [CSB+93], du GM-CSF [CSB+93] et de l'IL4 [RHG95] possèdent des modules NFAT/AP-1 et seule la fixation coopérative des protéines NFAT et AP-1 (telles que Fos et Jun) permet l'activation du promoteur sous-jacent. Il est intéressant de noter qu'une recherche systématique sur le génome humain de modules NFAT/AP-1 [KKMBW99] a montré que ces derniers étaient plus particulièrement retrouvés dans les promoteurs spécifiques des cellules T et ainsi suppose qu'ils jouent un rôle particulier dans les réponses immunitaires. Un autre exemple d'activation de la transcription grâce à la coopération de plusieurs facteurs de transcription est mis en évidence dans les séquences promotrices des gènes de l'IL8 [KLRS94] et de G-CSF [DCL+94] où la présence du module NF-κB/C/EBP est essentielle à l'expression de ces gènes. Il existe de nombreux autres exemples de coopération entre sites dans le cadre de l'expression des cytokines tels que NFAT et CRE sur le promoteur du TNFα [TJP+96].

Coopération et combinaison L'activation coordonnée des facteurs de transcription et de la fixation consécutive aux sites de liaison correspondants permet la formation de complexes multi-protéiques capables d'interagir et de recruter à la fois des co-activateurs et des composants de la machinerie transcriptionnelle. L'ensemble de tous ces acteurs réunis compose l'*enhanceosome*. Des exemples de coopération entre facteurs de transcription dans le cadre de l'expresssion des cytokines impliquent l'IFN-β, de l'IL6 et de GM-CSF. L'initiation de la transcription de l'IFN-β requiert l'activation des sites de liaison NF-κB, IRF et ATF2/c-Jun sur son promoteur, l'*enhanceosome* ainsi formé comprend les protéines IRF-3, IRF-7, NF-κB, p50, RelA et ATF2/c-Jun [ALP$^+$00]. La transcription du gène de l'IL6 nécessite la fixation sur son promoteur des protéines CBP/p300, CREB, AP-1, C/EBP et NF-κB [VBDBB$^+$99] alors que celle du gène de GM-CSF est issue de la coopération des facteurs de transcription NF-κB, Ets et AP-1 [TTS$^+$95]. L'arrangement spatial des sites des liaison sur le promoteur est très important et augmenter la distance séparant deux sites a pour effet de diminuer l'activation du gène.

Il existe une autre propriété qui influence la force d'activation d'un promoteur et permet de moduler à un niveau fin l'expression d'un gène. Différents *stimuli* peuvent enclencher la formation de complexes activateurs variés mais dont le résultat sur la transcription est similaire voire identique. C'est l'exemple du promoteur du TNF-α. Selon la nature des différents *stimuli* extérieurs et le type cellulaire impliqué, l'*enhanceosome* formé est différent. Dans le macrophage/monocyte, SP-1, Egr-1, Ets/Elk, ATF-2 et c-Jun ont été montré comme interagissant avec le promoteur de TNF-α afin d'activer sa transcription [TFT$^+$00]. En revanche, dans les lymphocytes B, d'une part Ets/Elk et SP1 semblent ne pas être recrutés sur le promoteur et d'autre part NFAT est ici nécessaire à l'initiation de la transcription du gène [TYTG96]. Dans les cellules T, la situation est un peu différente par rapport aux cellules B. NFAT a la possibilité de se fixer sur d'autres sites de liaison [TJP$^+$96] et selon le *stimulus* (par exemple, une infection virale) SP1 peut être nécessaire à l'activation du promoteur [FUB$^+$00].

1.4 Résumé

Dans ce chapitre nous avons introduit la notion de réseaux de régulation de gènes et l'importance de la combinatoire des facteurs de transcription dans l'expression des gènes.

De tels réseaux sont ubiquitaires en biologie. Connaître à la fois la connectivité et la

dynamique d'un réseau permet de comprendre les fonctions et les contraintes d'un tel système modélisé. De nombreuses techniques expérimentales récentes (par exemple, l'étude des profils d'expression, la découverte de motifs nucléiques, les expériences de *ChIP-on-chip* ou encore d'*ARN interférence*) et bioinformatiques commencent à être utilisées afin de cartographier et de disséquer les réseaux de régulation de l'expression des gènes. Néanmoins, ces méthodes ne sont pas les seules solutions envisagées afin de modéliser les réseaux biologiques. Une approche à la fois parallèle et complémentaire comprend l'utilisation des connaissances renfermées dans la littérature biomédicale. Les publications scientifiques représentent en effet une distillation des connaissances collectives du domaine qui peuvent être notamment exploitées afin de déterminer la connectivité d'un réseau de régulation de gènes, c'est à dire de répertorier l'ensemble des acteurs biologiques du système et leurs interactions.

Chapitre 2

La FdT pour la biologie

Le traitement automatique du langage naturel (TALN) est le traitement, par l'application de programmes et de techniques informatiques, du *langage naturel* (par exemple les langages humains), qui s'oppose au *langage de programmation*. Ces deux formes de langage diffèrent de manière fondamentale : l'interprétation d'un langage de programmation est de part sa conception non ambiguë alors que les interprétations possibles du langage naturel sont potentiellement ambiguës à tous les niveaux d'analyse. Le traitement des langages de programmation est un sujet pour les sciences de l'informatique et est classiquement enseigné dans des cours de développement de compilateurs. En revanche, le traitement des langues naturelles concerne de nombreuses disciplines, telles que la linguistique, les sciences de l'informatique et l'ingénierie. La fouille de texte (FdT), quant à elle, est d'une part la mise en œuvre de différentes méthodes de TALN et/ou de techniques purement statistiques afin d'extraire de l'information à partir de textes et d'autre part l'utilisation d'approches de fouille de données (FdD) dans le but de retrouver des associations parmi les données extraites de différents textes. Dans ce chapitre, il nous arrivera d'employer fréquemment le terme FdT dans un contexte plus restreint que celui proposé par la définition : l'aspect extraction de connaissances à partir de textes grâce aux techniques de TALN sera mis en avant au détriment des problématiques de FdD.

Un des développements les plus surprenants en bioinformatique a été l'attention toute particulière reçue par la FdT au sein des conférences de bioinformatique. Les conférences *PSB* (*Pacific Symposium on Biocomputing*) et *ISMB* (*Intelligent Systems for Molecular Biology*) ont commencé à publier des papiers sur le sujet dès le début des années 1990, allant jusqu'à leur dévouer des sessions entières à la fin des années 1990. Les communautés

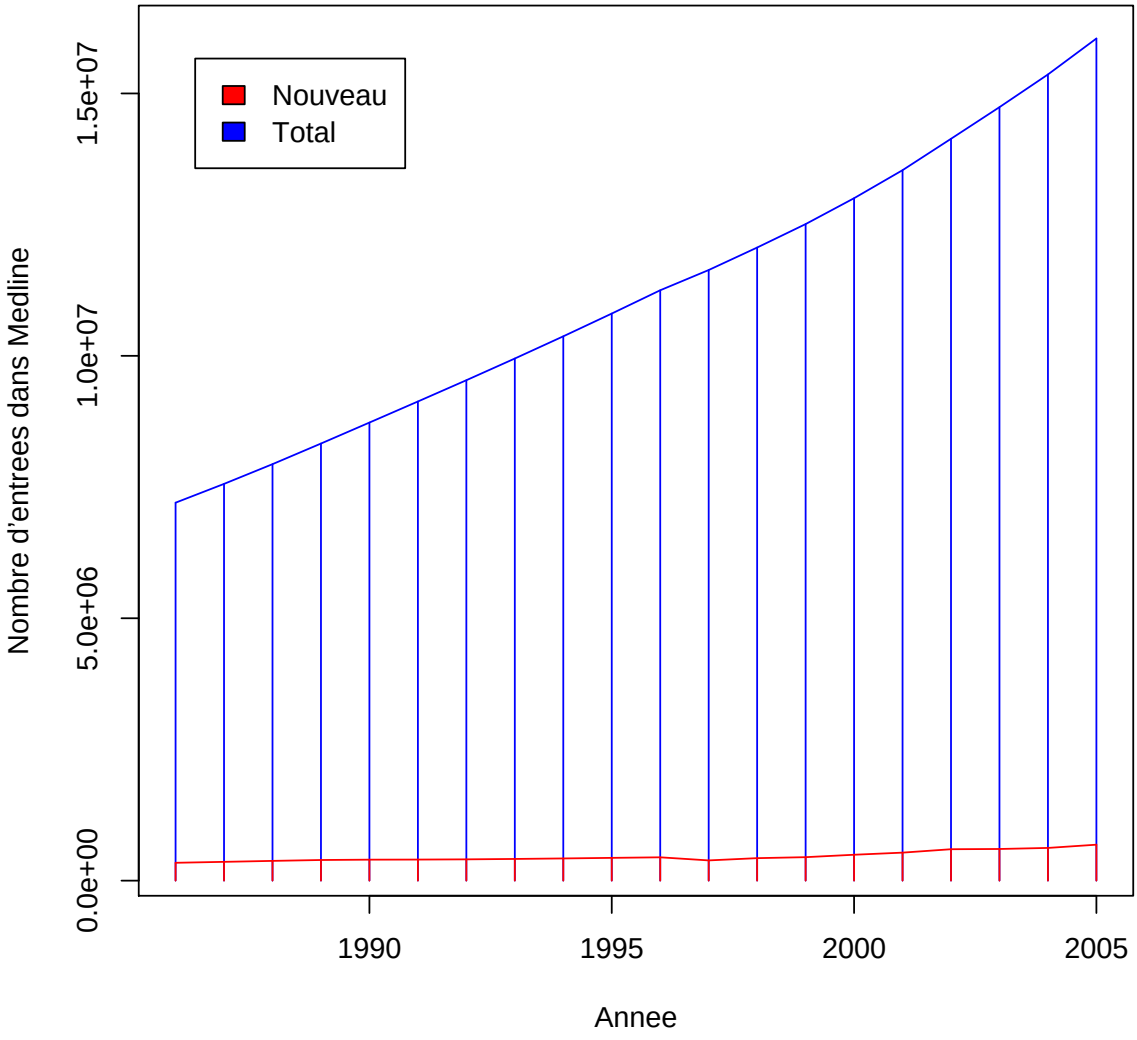
de la FdT leur ont rendu la pareille, notamment avec les conférences *ACL* (*Association for Computational Linguistics*) offrant récemment des ateliers de FdT appliqué aux domaines de la biologie. La rencontre des deux communautés à été bénéfique pour chacune, les biologistes disposant désormais d'outils leur permettant de nettoyer les ressources biomédicales accumulées depuis quelques années alors que les linguistes jouissent en retour de la mise à disposition de ces immenses ressources.

Le corps de littérature mis à disposition des biologistes est incroyablement imposant et grossit à une cadence effrénée. Même avec la mise à disposition des biologistes de structures, séquences et autres données expérimentales, la majeure partie de l'information biologique est encore sous la forme d'articles scientifiques [IEO01]. La base de données **Medline**[1] constitue la source bibliographique la plus complète à destination des acteurs du milieu biomédical, avec plus de 14 millions de publications scientifiques référencées en 2005. Voir la figure 2.1 pour référence. Il en résulte que l'analyse de son contenu ne peut plus être réalisée à l'échelle d'un seul individu [LHW03]. Parallèlement, l'information pertinente spécifique à un domaine particulier de la biologie se retrouve disséminé de plus en plus au travers des autres disciplines. L'explosion des expériences biologiques à très haut débit (le système double-hybride, le séquençage des *STS* (*Sequence Tagged Sites*) en génomique, les *EST* (*Expressed Sequence Tags*), la spectrométrie de masse en protéomique, les puces à ADN et les données d'expression, etc) se retrouve confrontée à la nécessité de réaliser la synthèse systématique de toute l'information publiée relative aux gènes et protéines étudiées. Derrière cette idée d'intégration de l'information, se cachent aussi les problèmes d'assimilation (recréer une histoire juste à partir de fragments de preuves disparates, dispersés un peu partout) et de dissémination (retrouver une information utile pour des personnes n'ayant pas connaissance de son existence). L'attention des biologistes moléculaires envers l'exploitation automatique de la littérature s'est accentuée et désormais cette tâche se révèle essentielle à leurs yeux.

Il est utile de souligner que l'ensemble des documents scientifiques répertorié par **Medline** sont écrits en langage naturel et plus particulièrement en anglais scientifique. Cette notion de langage naturel est à rapprocher de celle de texte non structuré. La compréhension de ce type de texte à l'aide de l'informatique repose sur l'utilisation de méthodes propres à la FdT. Les autres types de textes que sont les textes semi-structurés, que sont par exemple les phrases non grammaticales ou télégraphiques, ou les textes structurés, à base d'information *itémisée* (par exemple les tables d'un tableur, les documents XML),

[1] http ://www.pubmed.gov/

Fig. 2.1 – Croissance de **Medline**



La courbe bleue indique le nombre total d'entrées présentes à une date précise. La courbe rouge indique le nombre de nouvelles entrées spécifiques à l'année en cours.

sont lus par des moyens autres que ceux nécessitant des techniques de FdT à proprement parler. Malheureusement, la FdT est une discipline aussi difficile qu'indispensable. Elle utilise de grandes quantités de connaissances, et ce sur de très nombreux niveaux. Par exemple, le fait de savoir comment les mots sont formés (la morphologie) est cruciale pour comprendre des mots tels que *deubiquitinization* qui sont complexes et qui peuvent ne jamais avoir été vus auparavant. Le fait de savoir comment les expressions se combinent (la syntaxe) est nécessaire pour comprendre pourquoi une phrase telle que "These findings suggest that FAK functions in the regulation of cell migration and cell proliferation" est ambiguë (est-ce que FAK joue un rôle dans la prolifération cellulaire et dans la régulation de la migration cellulaire, ou joue-t-il un rôle dans la régulation de la prolifération cellulaire et dans la régulation de la migration cellulaire ?). Ces problèmes sont difficiles à résoudre - malgré le fait que depuis les années 1960 la plupart des acteurs de la FdT travaille sur la langue anglaise, l'analyse globale de la syntaxe anglaise est toujours hors de portée. Cependant, les difficultés majeures de la discipline demeurent liées à la représentation de la connaissance du monde que nous utilisons afin de comprendre le langage. En tant qu'utilisateur humain du langage, la connaissance du monde est à la fois tellement omniprésente et envahissante (et considérée comme allant de soi) dans notre compréhension du langage que nous en sommes généralement inconscients. Nous pouvons donc ignorer qu'elle joue en fait un rôle de premier plan. Si l'on considère les phrases "she boarded the plane with two suitcases" et "she boarded the plane with two engines" [JM03], chaque phrase est aussi ambiguë l'une que l'autre sur le plan syntaxique, avec deux constructions possibles pour chacune selon que l'on considère que c'est soit l'avion, soit la femme, qui a les valises ou les moteurs. Néanmoins, les humains ont peu de chance de retenir deux lectures différentes pour chaque phrase. Une seule analyse (et une différente pour chacune) semble plutôt évidente et exclusive. Ce phénomène est basé sur la connaissance que les humains ont à propos des avions et des êtres humains et des relation potentielles entre ces concepts. La représentation de ce niveau de connaissance, dans l'étendue et la profondeur, qui est nécessaire pour comprendre n'importe quel texte en anglais (ou dans tout autre langage) n'est pas réalisable à ce jour et dans un futur proche.

Dans cette section, nous présenterons brièvement les techniques basiques et les approches de FdT. Dans une première partie, nous mettrons en avant les différents niveaux d'analyse d'un texte et les modes de représentation de l'information extraite dans une approche de FdT. Dans une deuxième partie, nous nous focaliserons plus particulièrement sur les domaines d'application de la FdT pour la biologie. Nous détaillerons alors

les difficultés rencontrées dans les approches de FdT et les solutions proposées dans la littérature du domaine. Cette section sera replacée dans le contexte de l'extraction d'information pour la biologie et traitera successivement de ses deux composantes majeures que sont d'une part la reconnaissance des entités nommées et d'autre part l'identification des relations entre entités nommées.

2.1 L'extraction de connaissances à partir des textes

Il est nécessaire de définir ici quelques termes relatifs à la FdT. Ces différents termes correspondent à des applications particulières de la FdT qui peuvent être connectées entre elles ou intervenir à des moments distincts d'un même processus.

- L'EI (extraction d'information) analyse les textes écrits en langage naturel afin d'extraire des informations remarquables à partir de types pré-définis d'évènements, d'entités ou de relations entre entités (ou d'évènements). Ces données sont alors très souvent enregistrées dans des bases de données, automatiquement. Il n'est pas non plus rare d'analyser ces informations ainsi stockées pour ensuite en dégager des tendances, générer des résumés ou tout simplement pour les mettre à disposition en ligne. Voir la figure **??** pour illustration.
- La RI (recherche d'information) récupère, à partir d'une collection de documents textuels, un sous-ensemble qui est pertinent vis-à-vis d'une requête. Cette requête est très souvent à base de mots clefs, quelquefois en complément d'un thésaurus. Un exemple célèbre d'application en RI est le moteur de recherche **Google**. Parfois aussi la RI ne travaille pas sur un ensemble de documents mais sur un document unique. Dans ce cas le sous-ensemble pertinent qui est recouvré correspond à des portions de texte au sein de ce même document.

  L'EI n'est donc pas équivalent à la RI. Le premier récupère des faits à partir de textes, ce sont les faits qui sont analysés, alors que le deuxième récupère des collections de documents, on travaille ici à l'échelle du document.

Selon la tâche à effectuer, les applications de FdT doivent acquérir des connaissances de nature et de complexité variable à partir des textes. La figure 2.3 montre les différents niveaux d'analyse qui sont réalisables en EI. La difficulté est croissante de la base du schéma jusqu'au sommet. Les étapes situées au dessus du niveau du discours sont à ce jour encore irréalisables. Il est essentiel de noter que l'on ne peut procéder à l'étude des caractéristiques d'un niveau particulier sans avoir préalablement et correctement procéder à l'analyse du niveau directement inférieur. Il est très fréquent pour une application d'EI

Fig. 2.2 – Schéma de la procédure standard de EI

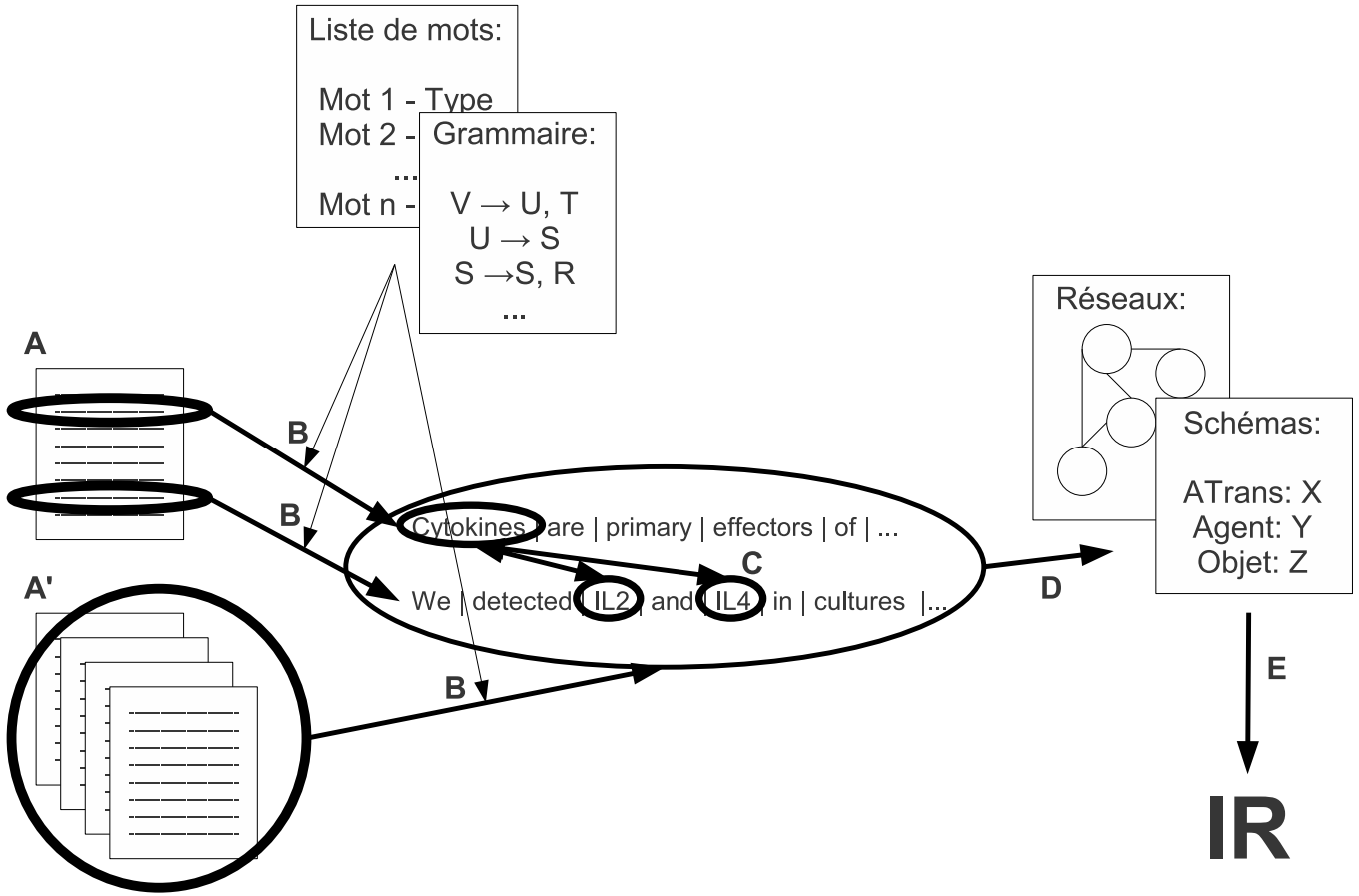


Schéma de la procédure standard de EI. Un ensemble de documents [A'] ou des portions de texte d'un document [A] sont analysés au niveau de la phrase [B] afin de retrouver des relations particulières entre les mots [C]. L'étape finale est l'intégration de l'ensemble de ces données [D].

d'aller jusqu'au niveau sémantique et même du discours. Classiquement, il existe 5 + 2 niveaux d'abstraction, classés dans l'ordre croissant :

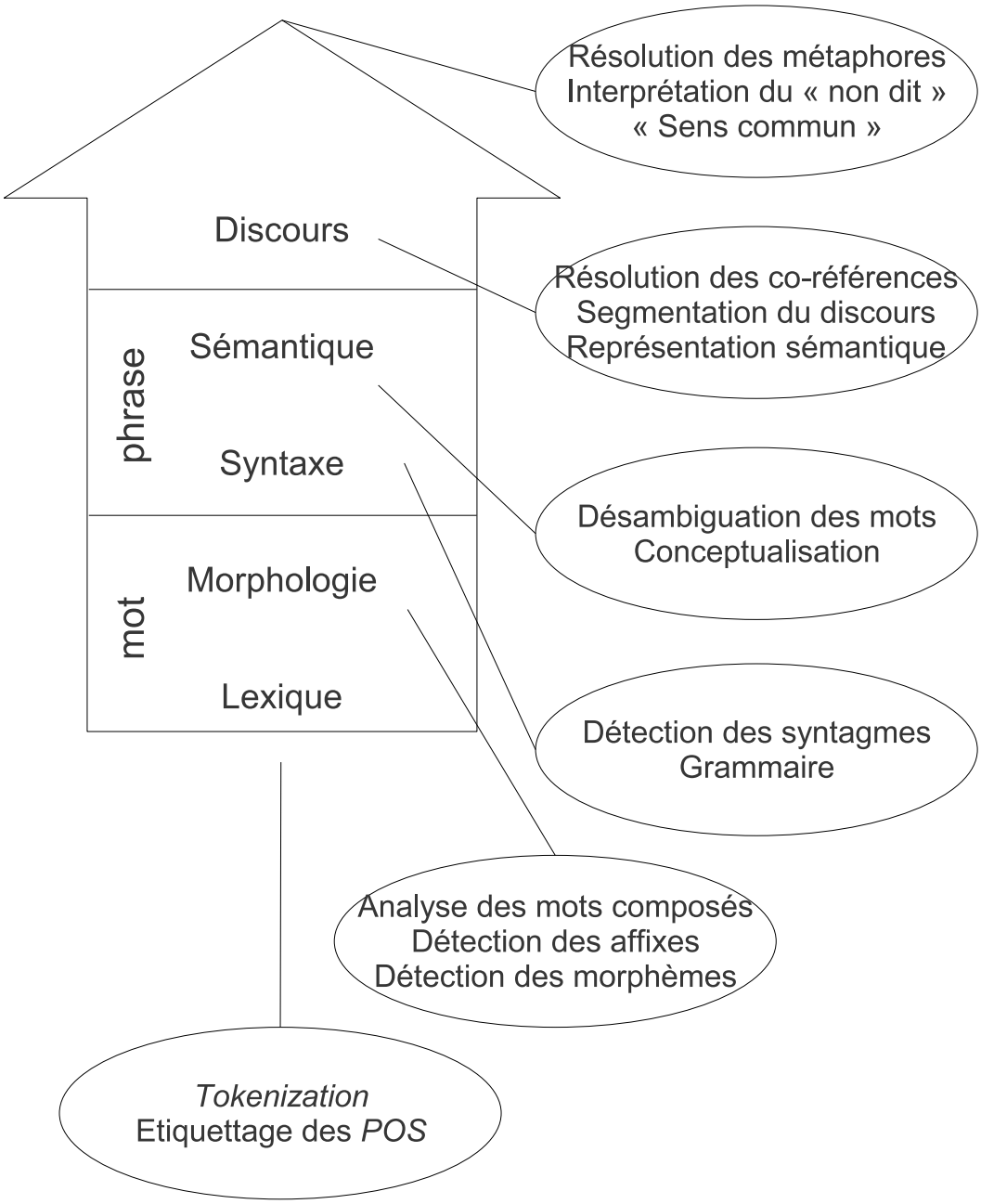
- le niveau lexical,
- le niveau morphologique,
- le niveau syntaxique,
- le niveau sémantique,
- les niveau du discours.

Le niveau phonologique, qui se situe en amont du niveau lexical, est spécifique des applications qui traitent de la synthèse vocal alors que le niveau le plus abstrait, celui de représentation du "monde réel", n'est pas envisageable au jour d'aujourd'hui ou dans un futur proche (en dehors de connaissances dans des domaines très spécifiques). Actuellement, les différentes approches de FdT décrites dans la littérature balayent le spectre complet des possibilités. Certaines utilisent de simples listes de mots (avec parfois leurs fréquences respectives d'occurrences, glanées à partir des documents eux mêmes) alors que d'autres prennent en compte des hiérarchies de concepts inter-connectés (exploitant par exemple les notions de généralisation, d'association ou d'agrégation). Ces méthodes sont alors considérées par leurs auteurs comme des réponses adéquates pour approcher au mieux les connaissances sur le "monde réel" implicitement contenues dans les documents à traiter. Ces solutions peuvent être indifféremment basées sur des caractéristiques purement statistiques des textes ou sur des représentations logiques selon les *a priori* des auteurs.

2.1.1 Le niveau lexical

Lorsqu'une représentation textuelle d'un document est accessible à l'outil informatique et disponible, une phase de *tokenization* (détermination les frontières des mots) est tout d'abord réalisée. Cette phase permet d'identifier les mots, délimiteurs, points d'arrêt, caractères spéciaux, etc présents dans le texte. Le rôle de l'analyse au niveau lexical est de regrouper les mots au sein de différentes catégories. Celles ci sont la plupart du temps des *POS* (cf l'annexe 6). Beaucoup d'approches de RI se focalisent uniquement sur des catégories particulières que sont les adjectifs, adverbes, noms et verbes. Elles contiennent en effet la sémantique du texte, en opposition aux autres *POS* qui sont des mots de structure uniquement (par exemple, les articles ou certains pronoms). La difficulté relative à l'identification des mots de la catégorie dite sémantique est que le nombre de mots la composant n'est absolument pas établi (par exemple, les noms des sociétés de

Fig. 2.3 – Le processus classique de FdT, décomposé



Les niveaux les plus abstraits ne sont pas nécessairement atteints selon l'application. Les niveaux lexicaux et morphologiques sont toujours mis en œuvre que ce soit en RI ou en EI (à quelques exceptions près).

maintenance informatique dans le monde, la liste des prénoms masculins, etc représentent un ensemble à la fois très grand et dynamique), à la différence des mots de la catégorie structurante qui peuvent aisément répertoriés car leur nombre est considéré comme fixé (ces mots sont définis par la grammaire de la langue). L'utilisation de dictionnaire n'est donc indiquée que pour identifier les mots de la classe dite structurante. Pour les mots de la classe sémantique, des techniques à base de règles ou d'heuristiques stochastiques (apprentissage à partir de modèles *n-grammes*) peuvent être employées pour assigner la catégorie la plus vraisemblable à un mot donné.

2.1.2 Le niveau morphologique

A l'étape morphologique, certaines simplifications lexicales peuvent être réalisées. Le but ici est de détecter autant d'apparitions d'un concept spécifique dans un texte que possible. Ces occurrences particulières pouvant faire croire que nous avons à faire à des concepts différents, masqués par des formes morphologiques distinctes, alors qu'il n'en est rien. Ce problème peut être considéré comme un forme particulière de co-référence. Notamment dans les approches statistiques, il est important de représenter correctement ces co-occurrences de concepts, souvent en utilisant les mêmes mots mais légèrement modifiés. Les co-références particulières que l'on peut résoudre à l'étape morphologique sont, d'une part la modularité des préfixes et des suffixes des noms, et d'autre part la conjugaison des verbes. La racinisation permet classiquement de répondre à ces difficultés.
Durant l'opération de racinisation, les différents temps d'un verbe sont réduits à leur racine, de façon à comparer plus facilement des relations dans un contexte donné. Généralement, l'algorithme simplifié se résume à :

```
Si finDuMot(mot) est ('ing' ou 's' ou 'ed') alors racine est mot
 sans finDuMot(mot)
```

Par exemple, si le mot est 'converts' sa racine est 'convert'. Les pluriels des noms sont passés au singulier pour les mêmes raisons. Par exemple, 'proteins' devient 'protein'.
Parallèlement, il est aussi nécessaire de retrouver les mots d'origine cachés derrière leurs dérivés. Typiquement, les affixes (grossièrement le préfixe et le suffixe d'un mot) sont caractéristiques des formes dérivées d'un mot. Un algorithme basique de simplification des formes dérivées peut être similaire à celui ci :

```
Si finDuMot(mot) est ('ly' ou 'ity') alors racine est mot
 sans finDuMot(mot)
Si debutDuMot(mot) est 'un' alors racine est mot sans debutDuMot(mot)
```

Le procédé de racinisation est relativement simple pour la langue anglaise et peut se résumer à un petit ensemble de règles seulement. L'algorithme de Porter [Por80] est l'exemple le plus connu d'algorithme de racinisation pour la transformation de noms, adjectifs et verbes. Celui-ci ne transforme pas les préfixes, uniquement les suffixes.
Un autre tâche réalisable à l'étape morphologique est la décomposition des mots composés. Néanmoins, les noms composés ne sont pas très fréquents en langue anglaise et des mots comme "database" ou "hardware", bien qu'à l'origine des mots composés, sont passés dans l'usage courant et ne sont plus décomposables. Ils expriment un concept unique né de la réunion des mots qui les composent.

2.1.3 Le niveau syntaxique

Au niveau syntaxique, les mots sont mis en relation les uns par rapport aux autres et permettent de lever des ambiguïtés qui ne pouvaient être résolues aux niveaux précédents. Deux tâches peuvent alors être réalisées : d'une part identifier les constituants ou fragments de la phrase (syntagmes) et d'autre part d'assigner de manière exacte le rôle de chacun des mots de la phrase. L'annexe 4.5 expose brièvement la théorie classique de la structuration du discours. Est prise alors en considération la grammaire du langage. A cette étape, on utilise un analyseur syntaxe dont le rôle est d'analyser la phrase en confrontant l'organisation des mots présents à une grammaire particulière. Cette grammaire est, rappelons le, un formalisme sur les structures de phrase permises au sein d'un langage. Des formalismes tels que les *grammaires de dépendances* permettent de préciser certains concepts et donc d'ajouter de la connaissance dans le texte. L'annexe 4.5 introduit succinctement la notion de grammaire et différents formalismes classiques.

2.1.4 Le niveau sémantique

Grossièrement, la connaissance acquise au niveau syntaxique inclus les catégories syntaxiques (telles que les noms, verbes, syntagmes nominaux) et comment ces catégories apparaissent ensemble dans les textes et sont ordonnées. De façon plus abstraite, la

connaissance syntaxique est une connaissance linguistique qui peut être établie sans aucune référence à ce dont les mots se réfèrent. La connaissance sémantique, elle, dépend des propriétés du sens des mots. Par exemple, la phrase "Interleukin 4 talks to the surgeon" est syntaxiquement correcte mais est sémantiquement incorrecte. L'Interleukine 4 est une protéine et n'est pas un être animé.
A ce stade, on souhaite ajouter ou compléter les concepts et les connaissances dans la représentation du texte. Le but ici est de comprendre le sens de la phrase en mettant en commun la représentation syntaxique du texte et le sens déjà assigné aux mots de manière individuelle. Il est aussi important que cette représentation du sens permette d'une part de raisonner facilement et qu'elle soit d'autre part assez riche pour lever les ambiguïtés sur l'interprétation de cette connaissance. Il existe donc deux tâches distinctes réalisables au niveau sémantique : la représentation de la connaissance et son extension mais encore la désambiguation du sens des mots.

Représentation de la connaissance Un état de l'art sur le problème de la représentation de la connaissance ne sera pas présenté dans ce document. Nous ne soulignerons ici que quelques solutions envisageables, les plus classiques dans leur forme. Elles sont brièvement exposées dans les paragraphes suivants, de manière croissante en terme de richesse de représentation apportée.

Représentation pauvre La forme la plus simple de représentation sémantique consiste à rester très proche de la grammaire d'origine tout en l'enrichissant. Ici, l'analyse syntaxique d'une phrase conduit à proposer plusieurs représentations (sémantiques) possibles de cette même phrase. Le calcul des prédicats est une logique formelle qui permet d'exprimer un grand nombre de raisonnements logiques. Il est souvent utilisé pour représenter le sens des phrases en langage naturel. L'avantage d'utiliser une telle technique est que son expression est non-ambiguë. Ainsi, un énoncé au sens ambigu en anglais correspond alors à plus d'une expression en logique des prédicats. Chaque représentation générée exprime une lecture particulière de l'énoncé d'origine. En revanche, calculer toutes les lectures d'un énoncé à partir d'une phrase très ambiguë se révèle extrêmement coûteux. Une alternative est d'exprimer l'énoncé en *formes quasi-logiques* [AM92] qui omet délibérément certains détails sur l'annotation de l'ambiguïté. Les lectures individuelles de l'ambiguïté ne sont pas réalisées à cette étape. En effet, énumérer toutes ces possibilités n'est pas toujours nécessaire à la résolution de l'ambiguïté et si cette dernière est d'ordre sémantique, et non d'ordre syntaxique, auquel cas la représentation se voit simplifiée (avec une seule dérivation dans la *forme quasi-logique*).

Représentation appronfondie Une forme de représentation sémantique un peu plus riche, est l'utilisation de *grammaires de cas*. Initiée par Charles Fillmore [Fil68], cette théorie propose d'analyser la phrase comme étant composée de verbes et de *cas* qui leurs sont associés. Un *cas* est ici défini comme une entité du texte non plus syntaxique mais sémantique (par exemple, l'Agent, le Lieu et l'Instrument). Un verbe est spécifique d'un certains nombre de *cas*, l'ensemble formant un *schéma de cas*. Ainsi, un *schéma de cas* décrit un aspect important de l'environnement d'un verbe, nom ou adjectif. L'hypothèse ici faite est que les fonctions grammaticales, comme le sujet ou l'objet d'un verbe, sont intimement liés aux *cas* présents dans la phrase. Ici la limite entre syntaxe et sémantique devient floue. Ainsi, lorsque Fillmore propose la hiérarchie universelle suivante pour sélectionner le sujet d'un verbe :

```
Agentive < Instrumental < Objective
```

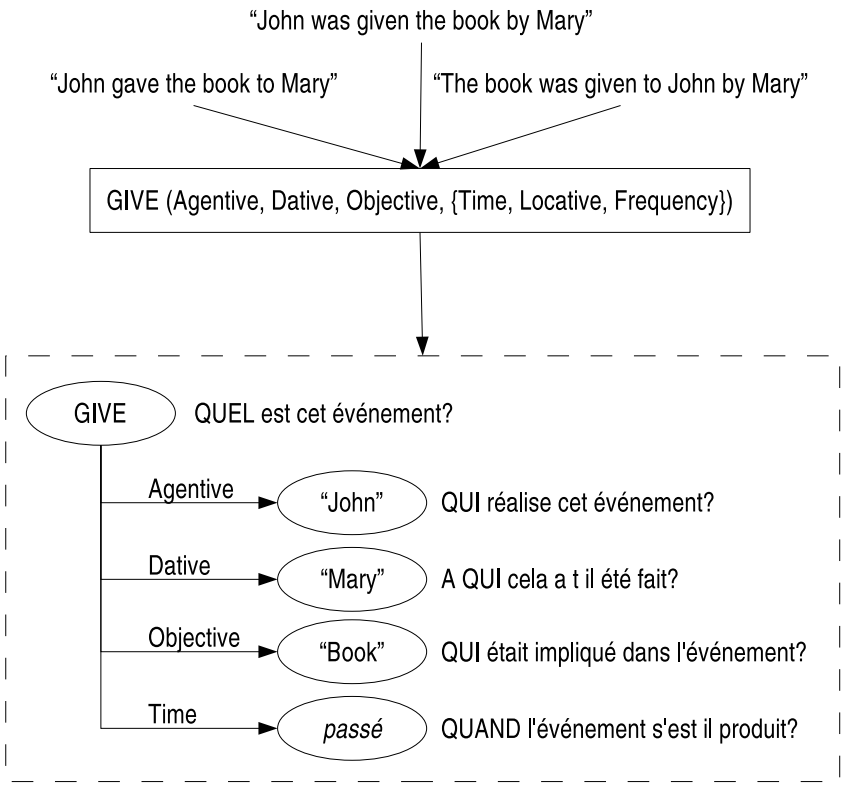
cela signifie que si le *schéma de cas* pour un verbe contient un Agent, celui-ci doit être le sujet du verbe à la voix active. Si il ne contient pas un Agent, alors c'est le *cas* qui suit dans cette hiérarchie qui est le sujet, soit l'Instrument, et ainsi de suite. Brièvement, les idées de base qui définissent une *grammaire de cas* sont exposés ci-dessous.

- La connaissance linguistique est organisée autour d'instances de sens particulier donnés aux verbes. Par exemple, dans les phrases "to run a gel shift" et "to run behind the bus", le sens du verbe 'run' est différent : dans la première phrase il décrit l'action de procéder à une expérimentation particulière de biologie moléculaire alors que dans la deuxième phrase il prend réellement la signification de courir.
- Il existe un nombre fini de *cas*. A titre d'exemple, il existe les *cas* :
  `Agentive: instigateur ou agent`
  `Instrumental: impliqué par une relation de causalité`
  `Datitive: l'entité affectée`
  `Factitive: le résultat`
  `Locative: le lieu`
  `etc`
- Chaque sens d'un verbe accepte un sous-ensemble particulier de ces *cas*. Ce sous-ensemble est alors partitionné en une fraction 'obligatoire' et une fraction 'optionnelle'. Un *cas* est obligatoire si son absence rend la phrase grammaticalement incorrecte. Par exemple, "John gave the book" est grammaticalement incorrect.

La figure 2.4 illustre un exemple de représentation sémantique avec une *grammaire de cas*. Néanmoins, de nombreuses critiques tendent à montrer que les *grammaires de cas* sont

insuffisantes pour représenter convenablement la sémantique [Nob88]. Il n'est pas rare que ce soit un nom qui convoie l'information (implicitement ou explicitement) et non un verbe. Par exemple, dans une phrase telle que "The binding of IL2 to its receptor is mandatory", le nom 'binding' implique l'idée que 'IL2' se lie physiquement à son récepteur. Dans ce cas, le *schéma de cas* doit aussi prendre en compte le sens donné aux noms.

FIG. 2.4 – Un exemple de représentation sémantique avec une *grammaire de cas*



Les trois phrases sont représentées sémantiquement de la même manière alors qu'elles sont syntaxiquement différentes.

Représentation riche Lorsque nous (êtres humains) lisons la phrase "John gave Mary a book" nous réalisons un certain nombre d'inférences, automatiquement et implicitement, à savoir que John a possédé le livre à un temps x, Mary a possédé le même livre à un autre temps y et que le moment x est antérieur au moment y. Ce type de représentation est à la base d'un extension majeure de la *grammaire de cas* et qui est souvent connue sous le nom de représentation à partir de *cadres* [Min75]. La figure 2.5 illustre cet exemple particulier. Cette approche permet aussi d'utiliser des phrases fortement "bruitées" ou incomplète. La phrase précédente, alors inutilisable par une *grammaire de cas*, "John gave the book", peut toutefois être informative dans le contexte ici. Nous ne travaillons plus uniquement

au niveau de la phrase mais du discours tout entier, si ce dernier est disponible. "John gave the book" peut être complétée par "The book is now Mary's", présente dans une autre phrase, afin de construire une véritable histoire. Un avantage non négligeable des *cadres* est alors de n'avoir à identifier que les principaux rôles et actions du discours, fournissant ainsi une représentation compacte de l'information contenue dans un texte imposant. Les *cadres* ont montré leur performance dans des domaines d'application très particuliers. Par exemple, le système *AQUA* [Ram87] est capable de recréer des récits complets à propos des incidents terroristes parus dans les dépêches de presse (Qui ? Quoi ? Comment ? Où ? etc). Pendant l'analyse syntaxique du texte, tous les indices qui peuvent se référer au contenu d'un *emplacement* (propriété atomique d'un *cadre*, par exemple une chaîne de caractères, un chiffre, d'autres *cadres*, etc) pour un *cadre* (instance d'une classe d'un objet ou d'un concept) spécifique servent à remplir cet *emplacement* particulier. La connaissance sémantique en relation avec une action particulière s'accumule au fur et à mesure de l'analyse du texte.

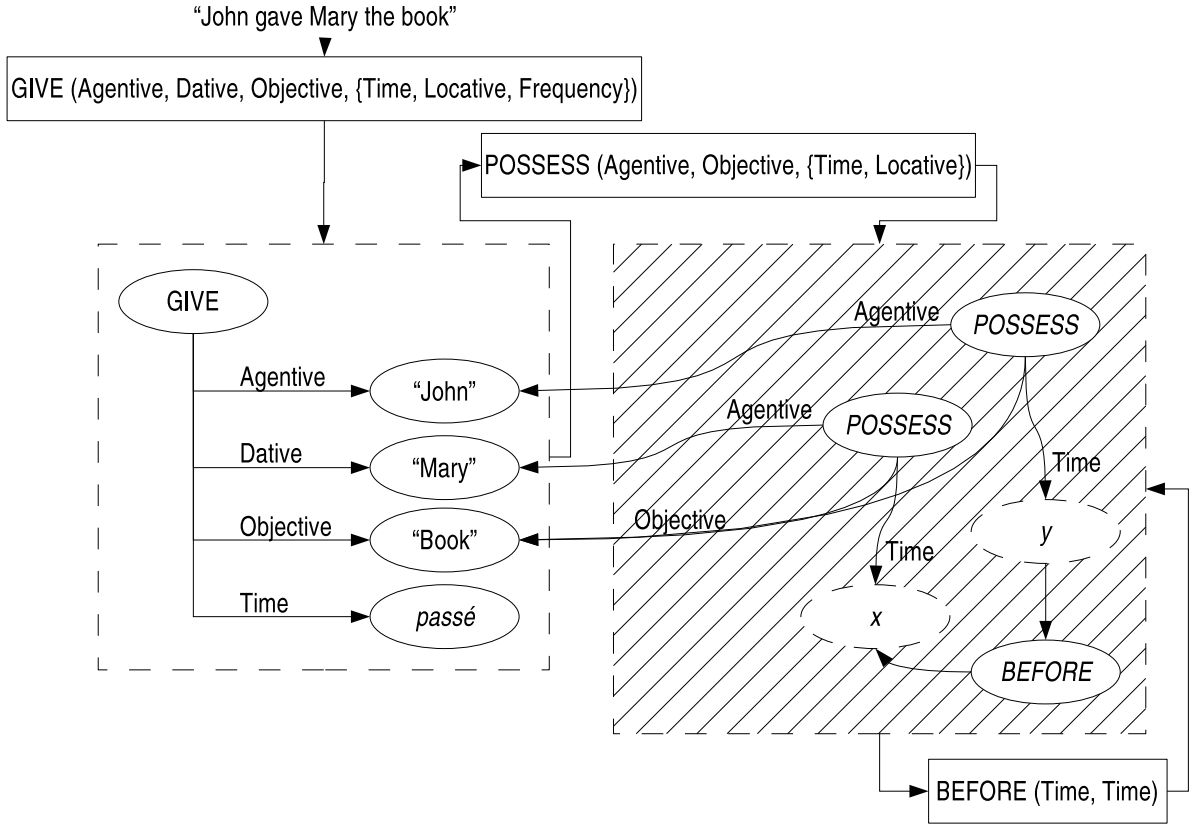
2.2 Les approches de FdT

Deux types d'approches s'opposent classiquement lors de la mise en œuvre d'un processus de FdT, en réalité elles sont souvent complémentaires et parfois utilisées de concert.

2.2.1 Approche basée sur les connaissances

Les grammaires sont construites manuellement et les motifs syntaxiques significatifs du domaine d'étude sont eux aussi découverts par des experts humains, grâce à l'étude et à l'inspection d'un corpus de textes. Cette approche se révèle être la plus fastidieuse, la mobilisation en personnel humain et en temps investi est généralement relativement important. Cela nécessite aussi de disposer de l'expertise adaptée. En revanche, et avec les compétences requises, le développement de systèmes de FdT extrêmement performants n'est pas conceptuellement compliqué. Les meilleurs systèmes développés à ce jour sont en majorité basées sur des approches d'ingénierie des connaissances.

Fig. 2.5 – Un exemple de représentation sémantique avec une approche à base de *cadres* (zone hachurée)



Les *cadres* utilisés sont **[POSSESS(John, Book, x)]**, **[POSSESS(Mary, Book, y)]** et **[BEFORE(x, y)]**.

2.2.2 Approche basée sur l'apprentissage automatique

Dès que possible, des méthodes statistiques sont mises en œuvre lorsque la notion de *réutilisabilité* est considérée comme importante. Les différentes règles nécessaires au déroulement du processus de FdT sont apprises à partir d'un corpus d'exemples annotés et/ou de l'interaction avec un utilisateur humain. Les méthodes d'acquisition de ces règles, centrées sur les données, permettent de couvrir l'intégralité des exemples. La portabilité d'un tel système d'un domaine d'étude à un autre est considérablement facilitée. De même, une expertise préalable dans le domaine d'étude n'est pas requise pour modifier et adapter ces systèmes. Néanmoins, les données d'exemples peuvent ne pas exister préalablement et être très difficile à générer. De même, en règle générale, la quantité de données demandée doit être largement supérieure par rapport à une approche d'ingénierie des connaissances pour des performances équivalentes.

2.3 Applications de FdT pour la biologie

L'importance de la FdT pour la biologie s'est accélérée en réponse à l'avènement des expérimentations biologiques à haut, voire à très haut débit. Des exemples d'applications de FdT aux données biologiques incluent la recherche automatisée dans la littérature d'ensembles de gènes impliqués dans une expérimentation, l'annotation de listes de gènes avec des concepts issus de *GO*, l'amélioration de la recherche d'homologie, la gestion des résultats de recherche dans la littérature, l'aide à l'entretien de bases de données et le peuplement de bases de données. Différentes approches et techniques de FdT ont été utilisées pour réaliser ces tâches, dont la bibliométrie et l'EI. Ces différentes approches peuvent être subdivisées en autant de sous-tâches. La réussite du processus dans sa globalité dépend de l'accomplissement de chacune de ces sous-tâches. Celles ci incluent l'extraction d'ENs, l'identification d'ENs, la *tokenization*, l'extraction de relations, l'indexation et la classification et catégorisation de contenu. Ces différentes tâches et sous-tâches seront présentées et explicitées dans les sections qui suivent.

2.3.1 Place de la FdT

La FdT s'insère dans l'analyse des données bioinformatiques de deux manières, ou plutôt à deux moments distincts. D'une part au début de la chaîne, en aidant à l'analyse des résultats générés par les expériences à haut débit, assistant ainsi les chercheurs à conduire un projet de l'expérimentation à la publication, et d'autre part, à la fin de la chaîne, en aidant le scientifique à exploiter le flux de publications issues de **Medline** (une moyenne de 1875 nouvelles entrées par jour en 2005). Dans une vue générale, la FdT peut aider le biologiste dans la réalisation de tâches à des échelles très variables : que ce soit, d'une part, celles limitées à une approche locale ou restreinte, comme localiser des données ponctuelles relatives à un seul ou un petit nombres d'objets biologiques (un gène, une protéine, etc) ou d'autre part, dans une approche systémique, de compiler des données à très grande échelle, comme par exemple peupler une base de données référençant toutes les interactions entre protéines chez *E. coli*.

2.3.2 Peupler et nettoyer les bases de données

Une des premières motivations des acteurs de la FdT en bioinformatique a été de remplir rapidement les bases de données d'informations d'intérêt en biologie. L'idée sous-jacente étant que si l'on peut détecter une référence à un objet biologique dans un texte, alors on peut remplir automatiquement les bases de données avec des faits relatifs à l'objet. Par exemple, les bases de données **BIND** [BBH03] et **DIP** [XRS+00, SMS+04] référencent les interactions inter-protéiques et utilisent des outils de la FdT. Le problème majeur contre lequel il faut faire face est le volume vertigineux d'information sur le sujet qui est présent dans la littérature biologique. Cette information doit aussi être transformée d'une forme non structurée donc, écrite en langage naturel, en une forme compréhensible par les outils informatiques, ici des tuples dans des tables de bases de données relationnelles. Le traitement peut alors être effectué *à la main*, en contrepartie de la mise en œuvre de ressources humaines et financières importantes ; l'utilisation de la FdT se limitant à extraire les portions de textes contenant l'information recherchée. Deux approches automatisées sont alors proposées pour éviter ce travail de longue haleine : d'une part la bibliométrie et d'autre part les méthodes d'EI. Les questions de peuplement de bases de données ont pour finalité de découvrir un nombre très restreint de catégories de faits - très souvent d'ailleurs une seule et unique catégorie de fait. Les approches bibliométriques sont basées sur le principe de co-occurrence simple d'objets biologiques d'intérêt. En reprenant notre exemple d'interactions protéine-protéine, cela signifie que si deux protéines sont mentionnées dans un même texte (que l'on se place à l'échelle de la phrase, du résumé ou de l'article en entier) alors on prend pour postulat qu'une relation particulière existe entre ces deux protéines. Les systèmes tels que **PubGene** [JLKH01] sont de bons exemples d'outils de bibliométrie pour la biologie. **PubGene** a la particularité de s'efforcer de s'affranchir du fait que deux protéines peuvent être présents dans le même texte par chance. La très grande majorité des applications de bibliométrie ne détectent que des paires de co-occurrence et en règle générale ces programmes souffrent de problèmes liés à l'extraction et l'identification d'ENs. Par exemple, **PubGene** ne peut identifier des références à des objets biologiques que grâce à leurs symboles et alias officiels qui ne peuvent s'étendre sur plus d'un mot. Il est aussi à noter que les outils actuels de bibliométrie génèrent un nombre incroyablement haut de faux positifs [YM02]. Les méthodes d'EI, plus fines, offrent une alternative moins souple mais peut être plus fidèle à la question de peuplement de bases de données [BACV99, CK99]. Les méthodes d'EI, de part le nombre plus élevé de contraintes conceptuelles impliquées, fournissent de manière générale de meilleures valeurs de précision que les méthodes de bibliométrie. Néanmoins, et à ce jour,

aucune méthode ne s'est avérée suffisamment fiable pour peupler automatiquement les bases de données sans nécessiter la validation d'un expert humain. Les bases de données majeures telles que **SwissProt** [BA00] ou **OMIM** [McK98] sont typiquement nettoyées et alimentées *à la main* à partir de la lecture de publications par des scientifiques expérimentés. L'utilité de la FdT peut alors ici venir par le biais d'une aide aux curateurs qui font face à ces impressionnants volumes de données. Récemment, bon nombre de compétitions de FdT ont été sponsorisées en ce sens par des organismes responsables de bases de données en biologie (à titre d'exemple, l'édition 2004 de la compétition *TREC Genomics* [Her05] a été supportée par le groupe *Mouse Genome Informatics*, responsable de la base de données **MGD** [EBK+05]).

2.3.3 Aide à l'analyse des expériences à (très) haut débit

De nombreuses études se sont spécifiquement penchées sur la question de l'analyse des données d'expression de gènes issues des expériences de puces à ADN. Un des systèmes pionniers en la matière, **MedMiner** [TSS+99], a été conçu pour extraire automatiquement à partir de publications les phrases mentionnant des relations entre deux gènes, un gène et une drogue ou faisant référence à des faits sur un gène unique. L'approche est ici adaptée à des ensembles de gènes volumineux qui sont retrouvés co-exprimés dans des expériences de puces à ADN. **MedMiner** permet de naviguer parmi des milliers d'articles retournés par le système en les classant par catégories d'intérêt pour les biologistes moléculaires. [SEWB00] décrit une méthode pour détecter les relations fonctionnelles dans les données de puces. D'autres approches [RCSA02] se sont focalisées sur l'amélioration de la classification des gènes répertoriés dans **GO** à partir de la littérature.

2.3.4 Interactions entre entités

Un autre corps d'étude de la FdT en biologie se concentre sur la découverte et la formalisation des réseaux d'interaction. Comme envisagé dans la section 1.1, la biologie peut être vue comme une science de réseaux : les interactions entre différentes entités biologiques (par exemple, les gènes, les protéines ou divers métabolites) à différents niveaux (par exemple, la régulation de gènes, la signalisation cellulaire) peuvent être représentés par des graphes. L'analyse de tels réseaux peut s'avérer utile pour proposer de nouvelles perspectives quant aux fonctions des systèmes biologiques. L'utilisation de réseaux biolo-

giques est d'une grande aide en bioinformatique : d'une part pour la modélisation et la simulation de systèmes biologiques [CDF+04], d'autre part en tant qu'aide à l'analyse de résultats expérimentaux [SEW05] et finalement la formulation de nouvelles hypothèses à partir de la mise en commun d'informations éparses et parfois implicites [SL04]. Différents types d'interactions et d'associations ont été étudiées grâce à la FdT, incluant les protéines entre elles [BACV99, BAOV99, OHTT01, DYE+04], notamment sur des critères fonctionnels [PKK01], les protéines et les composés pharmacologiques [RTWH00], les protéines et les pathologies ou localisations sub-cellulaires [CK99, SC03, SCR03], les réactions enzymatiques et des informations sur la structure protéique [HDG00]. En règle générale, les problèmes de FdT sont équivalents, quelque soit la nature exacte de l'interaction.

2.3.5 Indexation de la littérature

L'indexation de documents consiste à assigner des termes à des documents dans un contexte d'aide à la recherche et de récupération de documents. Cette tâche est particulièrement utile pour la gestion de bibliothèques et de bases de données. Le *National Library of Medicine*[2] indexe *à la main* **Medline** avec des termes sélectionnés à partir de **MeSH**[3], néanmoins un système automatisé a été développé en 2000 [ACH+00]. Ceci afin de pallier d'une part à l'augmentation des documents à indexer et d'autre part à la pénurie d'annotateurs compétents.

2.4 L'EI pour la biologie

Dans cette section, nous centrerons notre propos sur les problématiques liées à l'EI dans le domaine des textes de biologie. Nous présenterons deux étapes majeures de l'EI que sont la REN et la découverte des relations entre ENs. Un bref historique des travaux réalisés de même que les principales difficultés spécifiques à chaque tâche seront évoqués. En préambule, nous détaillerons les modalités de l'évaluation des résultats expérimentaux obtenus par les méthodes de FdT en biologie car les notions introduites seront employées tout au long de la section. Nous proposerons notamment de préciser les jeux de données couramment utilisés ainsi que les mesures de performance des systèmes.

[2] http ://www.nlm.nih.gov/
[3] http ://www.nlm.nih.gov/mesh/

2.4.1 Évaluation

2.4.1.1 Métriques

Les méthodes standards d'évaluation des techniques de FdT impliquent le calcul des valeurs de *précision*, *rappel* et parfois du *score F*. Ces mesures sont déterminées à partir des résultats obtenus de la FdT que l'on compare à des résultats étalon. Les données étalon sont, dans le meilleur des cas, des données annotées par des experts humains et révèlent alors les performances maximales qu'un système de FdT peut espérer atteindre. La préparation de telles données étalon représente une très grande gageure pour les scientifiques de la communauté de biologie, que se soit en terme d'attente des acteurs de la FdT et de la qualité des annotations.

La *précision* mesure le taux de résultats individuels pertinents produits par le système. Cette valeur est similaire à la notion de spécificité et s'établit de la manière suivante : $P = \frac{VP}{VP+FP}$ où VP est le nombre de vrais positifs (le nombre de résultats individuels corrects) et FP le nombre de faux positifs (le nombre de résultats individuels incorrects). Le *rappel* quant à lui mesure le taux de tous les résultats individuels pertinents trouvés par le système. Le *rappel* est similaire à la sensibilité et se calcule de la façon suivante : $R = \frac{VP}{VP+FN}$ où FN est le nombre de faux négatifs (le nombre de résultats individuels que le système aurait dû compter comme bons mais ne l'a pas fait). Le *score F* standard, ou moyenne harmonique, tente de réaliser la balance des contributions de la *précision* et du *rappel* dans les performances globales du système. Il est défini par : $F = 2\frac{P \times R}{P+R}$. D'autres métriques sont parfois utilisées en supplément ou plus rarement en remplacement des mesures de *rappel*, *précision* et de *score F* [JM03]. Les différentes mesures d'évaluation des résultats exposées ici proviennent de la communauté de la RI. Dans cette approche l'unité de résultat est le document. Une tâche classique en RI revient à retrouver les documents pertinents parmi un ensemble de documents. La notion de pertinence est liée à la réponse attendue à une requête spécifique (par exemple, "quels sont les documents qui traitent du métabolisme du rat parmi tous les documents indexés par **Medline** ?"). Les documents pertinents et détectés en ce sens par le système correspondent à l'ensemble des vrais positifs. Les documents détectés par le système en tant que documents pertinents mais qui ne le sont pas sont alors les faux positifs. Finalement, les documents pertinents que le système n'a pas su détecter alors qu'il l'étaient représentent la population des faux négatifs.

2.4.1.2 Evaluation

La plupart des systèmes décrits dans la littérature ont été évaluées à partir d'étalons préparés par les auteurs eux-mêmes et non librement diffusés. Néanmoins, et ce depuis quelques années désormais, les premières "compétitions" de FdT en biologie ont vu le jour. Ces compétitions proposent à chaque groupe participant d'être évalué par une entité indépendante sur les mêmes données étalon. Les compétitions du domaine les plus réputées étant (sans ordre particulier) :

- Le *KDD Cup Genomics Challenge*[4]
- Le *TREC Genomics Track*[5]
- *BioCreative*[6]

Quelques jeux de données de test couramment utilisés seront présentés dans les sections relatives à la présentation de la REN et de l'EI en biologie.

2.4.2 REN

Toutes les approches de FdT en bioinformatique nécessitent de reconnaitre les références faites aux différents objets manipulés dans les textes. Ces objets étant très généralement les gènes ou les protéines, la plupart des applications s'adressant en effet aux communautés de la biologie moléculaire. Par exemple, dans la phrase "Thus, endogenous IL-2 is important in regulating expression of p27 as well as p21 (...)", IL-2, p27 et p21 sont des noms d'objets biologiques et plus particulièrement le premier est le nom d'une protéine et les deux derniers les noms de deux gènes. La tâche qui consiste à reconnaitre les apparitions de noms d'objets et à pouvoir les classer selon leur nature à partir des textes rédigés en langage naturel est connu sous le nom générique d'extraction d'entités nommées (EEN). Ce problème n'est pas spécifique aux approches de FdT en bioinformatique et cette dénomination date probablement des toutes premières conférences *MUC*[7], conférences historiques s'adressant des problèmes généraux de la FdT en langue anglaise [Sun93]. Dans la langue anglaise courante, les ENs sont assez hétérogènes ; allant des noms propres aux dates, en passant par des sommes d'argent. Dans les domaines spécialisés et techniques l'ensemble des ENs peut être plus restreint. En revanche le niveau de granulosité demandé dans la segmentation des classes d'ENs est souvent plus fin. Il est intéressant

[4] http ://www.kdd2006.com/
[5] http ://ir.ohsu.edu/genomics/
[6] http ://www.mitre.org/public/biocreative/
[7] http ://www.muc.saic.com/

de noter que si la REN appliquée à des domaines généraux semble atteindre des niveaux de performance comparables aux annotateurs humains [BSAG98], le problème est tout autre dans les domaines techniques et notamment en biologie moléculaire. Les approches rapportées d'EEN dans la littérature du domaine sont classiquement de deux types : d'une part les approches à base de règles et d'autre part les approches basées sur des techniques d'apprentissage automatique. Néanmoins, de nombreuses solutions mixtes ont aussi vu le jour : l'apprentissage automatique peut ainsi servir à inférer des règles (par exemple [Hea92]). Les approches à base de règles utilisent très généralement des combinaisons d'expressions régulières pour définir des motifs (ou patrons) qui peuvent correspondre à des noms d'entités dans les textes. Il est alors courant d'utiliser en parallèle une logique particulière pour permettre l'extension du nom de l'entité à droite et à gauche du patron et ainsi cerner les limites exactes du nom de l'entité. A titre d'exemple, l'expression régulière

```
[a-zA-Z]+[\-]?[0-9]+
```

(une séquence de lettres en majuscule ou minuscule suivie ou non par un tiret et terminée par un nombre quelconque de chiffres) reconnaitra les instances IL-2, p27 et p21 de notre exemple précédent (en début de paragraphe) en tant qu'ENs. Le travail de [FTTT98][8] dans ce domaine a été précurseur de bon nombre d'applications d'EEN en biologie à base de règles. En parallèle aux méthodes à base de règles, une grande variété de techniques d'apprentissage automatique ont été essayées. Celles-ci incluent les arbres de décisions, les classifieurs de Bayes naïfs, les *HMM* et les *SVM*. La très grande majorité de ces approches s'attache à détecter la localisation d'une EN au sein des textes. En revanche, peu d'entre elles permettent de faire correspondre une instance d'EN à des objets biologiques répertoriés dans des bases de données. Cette tâche est alors appelée identification des entités nommées (IEN) et se situe en aval de l'EEN. L'ensemble des deux tâches, EEN et IEN, constituant la tâche globale de REN. L'EEN, utilisée de manière isolée, est particulièrement adaptée à la découverte de nouveaux noms d'entités biologiques. En revanche, l'IEN s'avère être essentielle à l'identification des instances particulières d'ENs détectées. L'IEN repose sur l'utilisation de dictionnaires d'ENs et de techniques qui permettent de rechercher des termes de dictionnaires dans les textes. Historiquement, les premiers systèmes d'IEN ont cherché à mettre en correspondance directe les entrées d'une base de données de gènes ou de protéines avec la sortie obtenue à l'étape d'EEN. Il se révèle que ces systèmes obtiennent des résultats assez médiocres. En effet, très peu de noms de gènes ou

[8] http ://www.hgc.jp/service/tooldoc/KeX/intro.html

de protéines peuvent être directement retrouvés dans les textes à partir des appellations présentes dans les bases de données. Il en résulte que des dictionnaires adaptés doivent être construits pour répondre à ce problème. La qualité de construction de ces dictionnaires est critique et joue un rôle majeur dans les performances de l'IEN. Deux méthodes d'élaboration de dictionnaire s'opposent : la création manuelle de dictionnaire [OHTT01] ou semi-automatique [HFMZ03]. Dans les méthodes semi-automatisées, et généralement, une procédure semi-automatique permet de générer et de nettoyer un dictionnaire à partir du contenu de bases de données généralistes et publiques.

2.4.2.1 Historique

Fukuda *et al.* [FTTT98] ont été des pionniers dans la REN en biologie. Dans leur premier papier, ils ont montré la capacité de leurs méthodes à détecter des noms de protéines. Ils sont partis de trois constatations très simples et spécifiques aux articles biomédicaux et ont développé des techniques adaptées à ces cas particuliers. D'une part, les noms de protéines sont souvent des noms composés très long. D'autre part, des noms différents sont utilisés pour identifier la même protéine. Finalement, certains noms de protéines sont aussi des mots de l'anglais usuel. La solution apportée à ces difficultés est l'utilisation de caractéristiques spéciales telles que la présence de lettres majuscules, de chiffres et de terminaisons spéciales dans les mots décrivant les protéines. Les auteurs n'ont pas présentés de valeur de précision afin de valider leurs méthodes ; de nombreuses approches similaires ont été proposées ultérieurement et ont suggérées qu'une précision et un rappel supérieurs à 70% pouvaient être atteints. La création d'un large corpus de biologie par Otha *et al.* [OTC+00] un an plus tard, ainsi que le développement de techniques basées sur les Modèles de Markov Cachés [CNT00] ou sur des classifieurs Bayésiens entraînés sur des n-grammes [WHD+99] ont permis d'augmenter encore le rappel et la précision pour la reconnaissance de noms de protéines. Plus tard, la REN dans le domaine biomédical ne s'est plus cantonnée aux seuls noms de protéines et s'est diversifiée. Narayanaswamy *et al.* [NRVS03] proposent une extension au système développé par Fukuda *et al.* et peuvent détecter des noms de composés chimiques, de groupements ou radicaux chimiques, de types cellulaires ou d'organismes en plus des noms de gènes ou protéines. Les auteurs reprennent la notion de *termes cœur* et de *termes fonctionnels* de Fukuda *et al.* et précisent que la classe effective d'une entité est déterminée à l'aide des *termes fonctionnels*. Par exemple, dans le texte "CREB binding protein", "CREB" est un *terme cœur* alors que "protein" est un *terme fonctionnel* et spécifie que l'entité appartient à la catégorie

biologiques des protéines/gènes/ARNs. Frantzi *et al.* [FAM00] utilisent une méthode combinant linguistique et statistique afin de détecter les bornes d'une EN dans le texte. Cette méthode n'est pas limitée a des classes particulières de concepts. Elle a été plus tard perfectionnée par Nenadic *et al.* [NSA03]. Des termes candidats sont tout d'abord extraits des phrases en utilisant un ensemble de filtres linguistiques. Ces termes candidats représentent alors potentiellement des noms d'ENs. Un score de vraisemblance d'appartenance à un nom d'entité nommé est associé à chaque terme extrait durant l'étape précédente. Cette mesure, statistique, combine quatre caractéristiques numériques apprises sur un corpus d'apprentissage. Les quatres caractéristiques retenues sont :

1. la fréquence d'occurrence du terme dans l'ensemble des textes d'apprentissage,
2. la fréquence d'occurrence du terme au sein des chaines de caractères qui composent les autres termes,
3. le nombre de termes candidats inclus dans d'autres termes,
4. le nombre de mots composant le terme candidat.

Un raffinement de la méthode est apporté en prenant en compte le contexte local des termes candidats. Ainsi les mots jouxtant les termes candidats peuvent être intégrés à ces derniers s'ils sont généralement retrouvés de pair. Nenadic *et al.* proposent alors de grouper les ENs détectées en fonction du rôle sémantique sous-jacent de chacune et ainsi de structurer la terminologie extraite. Deux métriques sont principalement utilisées : d'une part la similarité lexicale et d'autre part la similarité syntaxique. La similarité lexicale se base sur la présence commune de *têtes* ou de *modificateurs* entre deux ENs. Par exemple, "Progesterone Receptor" et "Estrogen Receptor" partagent la même *tête* et peuvent indiquer que les deux ENs partagent un concept commun hyperonyme. Parallèlement, la présence d'un *modificateur* peut suggérer au contraire un lien de spécialisation entre deux ENs. Par exemple, "Orphan Nuclear Receptor" et "Nuclear Receptor". La similarité syntaxique essaie de mesurer le *parallélisme* entre deux constructions syntaxiques telles que l'énumération, la coordination, l'apposition et l'anaphore. L'hypothèse établie par les auteurs prévoit que deux ENs syntaxiquement similaires sont aussi *fonctionnellement* similaires. Par exemple, dans la phrase "Transactivation by either estrogen receptor or progesterone receptor involves a conserved AF-2 domain", "Progesterone Receptor" et "Estrogen Receptor" sont supposés partager une fonction similaire. Les auteurs obtiennent sur leur corpus de test une précision de 99% et un rappel de 74%.

Dans les travaux d'IEN de Hanisch *et al.* [HFMZ03], un dictionnaire spécifique aux noms de protéines est élaboré à partir de la base de données de nomenclature **HUGO** [WLD+04]. Chaque entrée du dictionnaire contient ainsi le nom complet et de possibles

alias de symboles **HUGO** de gènes. Le dictionnaire est complété avec des synonymes issus des bases de données **OMIM**[9], **Swiss-Prot** et **TrEMBL**[10]. Le nettoyage du dictionnaire inclut une phase d'extension qui propose de développer les acronymes non ambigus sous leurs formes complètes, issues du dictionnaire. Par exemple, l'acronyme IL2 est développé en Interleukin 2. Une deuxième étape de nettoyage, dîtes d' "élagage", propose de supprimer les entrées redondantes, ambiguës et erronées. Lorsque le dictionnaire est construit, les auteurs proposent de décomposer les phrases des documents en unités de mots. Ne sont conservés que les mots qui appartiennent à des mots du dictionnaire. A chaque mot est associé deux scores, un score d'*acceptation* et un score de *détection de bordure*. Le score de *détection de bordure* contrôle la fin de l'extension d'un mot appartenant potentiellement à un nom de protéine et représente la probabilité que le mot n'appartienne pas à un nom de protéine du dictionnaire. Le score d'*acceptation* détermine lui la vraisemblance du mot à appartenir à un nom de protéine issu du dictionnaire. L'utilisation de ces deux scores permet aux auteurs d'identifier des instances de protéines dans les textes. **MetaMap Transfer (MMtX)** [Aro01] utilise le **Metathesaurus UMLS** afin d'identifier les ENs. Les groupes nominaux sont tout d'abord isolés des phrases puis les *têtes* sont séparés des *modificateurs*. Dans un deuxième temps, pour chaque expression d'un groupe nominal extrait, des variantes sont générées. Une variante est ici soit un acronyme, une abréviation, un synonyme, un dérivé orthographique, un lemme ou toute combinaison autorisée de ces différentes formes. Les auteurs utilisent un lexique spécialisé afin de générer les acronymes, les synonymes et les abréviations. Un score est alors donné à chaque variante en fonction du nombre de transformations effectuées, plus le score est fort et plus la variante a subi de transformations. Un ensemble d'ENs candidates est rapatrié du **Metathesaurus**, un candidat doit contenir au minimum une variante commune avec le groupe nominal extrait et transformé. Finalement, chaque candidat est évalué individuellement face à ce dernier. Sont pris en compte dans l'évaluation les mots le composant en fonction de quatre critères : l'implication ou non d'une *tête*, la moyenne des scores de transformation, combien de mots sont retrouvés en commun et le nombre de caractères identiques entre les mots partagés. Une liste de candidats est ainsi générée et ordonnée pour chaque EN potentielle, un filtrage basé sur la mesure d'évaluation obtenue permet de ne garder que les candidats les plus vraisemblables. Aucun module de désambigüisation n'est proposé. **AZuRE** [PCG+04] utilise des techniques d'apprentissage par la machine dans un contexte d'IEN pour les gènes et les protéines. La méthode repose sur la génération de modèles gène spécifique versus tous les autres gènes. Un modèle est ainsi créé pour chaque gène réfé-

[9] http ://www.ncbi.nlm.nih.gov/sites/entrez ?db=omim
[10] http ://us.expasy.org/sprot/

rencé dans **LocusLink** ce qui correspond à environ 20000 modèles différents à la date de publication de l'article décrivant ces travaux. Les paramètres de modèle sélectionnés sont les distributions des mots associés à l'entité dans les textes de description de **LocusLink**.

2.4.2.2 Difficultés récurrentes communes aux textes de différents domaines

Variations des termes Cette difficulté est à rapprocher du problème plus global de la synonymie. L'ambigüité est néanmoins ici beaucoup plus restreinte et ne concerne que des divergences dans la *façon* d'écrire et non de l'utilisation de mots lexicalement différents. Les différentes formes de variabilités de termes sont d'ordre

- orthographique. La ponctuation (par exemple, "all-trans-retinoic acid" et "all trans retinoic acid"), la casse de caractères (par exemple, "ZFH2" et "Zfh2") ou l'utilisation de caractères et symboles étrangers (par exemple, "IL2 beta" et "IL2 β ou "Zfh II" et "Zfh 2") peuvent être transparents et être mélangés lors de l'écriture. En revanche, et dans certains cas, ces informations de type orthographiques permettent de différencier deux entités distinctes (par exemple, "cAMP" et "CAMP" sont deux objets biologiques différents). Les conventions d'écriture dans les documents techniques sont extrêmement importantes alors qu'elles sont généralement plus permissives dans les domaines plus généraux.
- Morphologique. Ce problème est en général lié à l'utilisation de suffixes ou de préfixes différents. Les différences observées peuvent être dues à la langue (par exemple le suffixe américain *-or* et *-our*, britannique, sont équivalents) ou fixées par des contraintes grammaticales (par exemple "biochemical study" et "biochemical studies" convoient la même sémantique, à la différence de "regulated protein" et "regulating protein").
- Syntaxique. Ces différences peuvent apparaitre dans la structure de termes composés par plusieurs mots. Par exemple, "human clones" et "clones of human" sont synonymes.
- Ou encore lexico-semantique. Des mots d'un même champ lexical peuvent être interchangés. Par exemple, "eye surgery" et "ophtalmologic surgery" sont synonymes.

Différentes méthodes ont été développées dans le but de gérer les différentes variantes de termes. Certaines méthodes sont spécifiques au domaine de la FdT sur les textes de spécialité biomédicales. Krauthammer *et al.* [KRMF00] ont ainsi mis au point un système basé sur **BLAST** afin de reconnaitre les variantes orthographiques de noms de gènes et de protéines et utilisant des techniques de mise en correspondance approximative de

portion de textes et de contenu de dictionnaires. **FASTR** [Jac01] utilise des méta-règles pour décrire la normalisation des termes et gérer à la fois les variations morphologiques et syntaxiques. Les variations sémantiques sont elles prises en charge grâce au recours de **WORDNET**[11].

Constructions imbriquées L'utilisation de conjonctions et de disjonctions complique l'association *a posteriori* de *têtes* ou de *modificateurs* à chaque membre de la construction. Par exemple, dans la construction "Human B- and T-cell lines", deux solutions possibles au développement de la conjonction sont d'une part, "Human B-cell lines" et "T-cell lines", et d'autre part, "Human B-cell lines" et "Human T-cell lines". Une étude menée sur le corpus **GENIA**[NCT00] montre que 5% de l'ensemble des termes spécifiques à la biologie moléculaire sont impliqués dans une constructive soit conjonctive soit disjonctive.

Co-références et anaphores Cette difficulté est liée à l'utilisation dans une phrase d'un mot ou d'un groupe de mots qui se réfère à une entité introduite plus tôt dans le discours [Mit99]. Dans ce document nous différencions les anaphores (par exemple, "*Il-2* is a regulatory protein. Its effects have been studied for several years now."), qui se manifestent sous la forme de pronoms ou du mot 'one' (l'utilisation du pronom singulier indéfini, 'on' en français), et les co-références (par exemple, "*Il-2* is a regulatory protein. This protein effects have been studied for several years now.") qui elles sont présentes sous la forme de groupes nominaux.

Les approches classiques [Mit99] utilisent une combinaison de *facteurs* afin de résoudre les anaphores ou les co-références. Ces *facteurs* peuvent tout aussi bien aider à éliminer des solutions peu vraisemblables ou au contraire donner une préférence à certains candidats plutôt que d'autres. Les *facteurs* discrimants négatifs sont classiquement :

- les discordances du genre et du nombre entre l'anaphore ou la co-référence et la solution candidate.
- Le non-respect de contraintes établies par des théories syntaxiques gouvernant les relations entre les nœuds d'un arbre syntaxique (voir les travaux d'Ingria *et al.* [IS89] pour une description de ces théories et une application à la résolution des anaphores).
- Le décalage sémantique entre l'anaphore ou la co-référence et la solution candidate. En étudiant le contexte d'apparition de l'anaphore ou de la co-référence, il s'agit de

[11] http ://wordnet.princeton.edu/

juger de la compatibilité des actions subies ou déclenchées entre les deux protagonistes. Par exemple, dans la phrase "We first injected *the gene* in *the cell* and then articially expressed it" la solution à l'anaphore 'it' ne peut vraisemblablement être "the cell' car une cellule ne peut s'exprimer dans un contexte de biologie moléculaire.

Par ailleurs, les facteurs discriminants positifs sont classiquement :

- le respect du parallélisme syntaxique. Les solutions dont les fonctions syntaxiques sont identiques à celle de l'anaphore ou de la co-référence sont favorisées. Par exemple, dans la phrase "*The gene* has been activated within *the articial promoter construct*, but it has failed in our first experiment" 'The gene' peut être privilégié à 'the articial promoter construct' en tant que solution de l'anaphore 'it' car ils partagent la même fonction de sujet dans la phrase.
- L'adéquation avec le *thème* principal du contexte. Lorsque plusieurs solutions restent en lice à la fin du processus de résolution des anaphores ou des co-références, il peut être intéressant de cerner la solution la plus mise en avant dans le texte de part le *thème* (ou centre d'attention tout au long du discours) de la section du document concernée. Par exemple, aucun système automatisé ou lecteur humain n'est capable de résoudre l'anaphore "We co-cultured cells A along with cells B and analyzed its evolution". En revanche si l'on se place à l'échelle du discours et que l'on remarque que le sujet central du document est 'cell A' au détriment de 'cell B' alors il est envisageable de préférer 'cell A' en tant que solution de l'anaphore.

De nombreux algorithmes de résolution des anaphores et des co-références ont vu le jour dans le domaine des textes généralistes et combinant tout ou partie de ces *facteurs* à partir des années 80. Par exemple nous pouvons citer les travaux de Lappin *et al.* [LL94], Kennedy *et al.* [KB96] ou encore le logiciel **CogNIAC** [Bal97]. Les premiers travaux traitant des problèmes de co-références et d'anaphores sur les textes de biologie sont probablement ceux de Pustejovsky *et al.* [PCZ+02]. Ils ont développé un analyseur robuste afin d'identifier et d'extraire les relations d'inhibition unissant différentes ENs d'intérêt. Leur système de résolution d'anaphores est basé sur les méthodes de Kennedy *et al.* et appliqué au domaine de la biologie. Sur 10 anaphores ou co-références repérées dans leur corpus de test, 8 ont été correctement résolues.

Autres liens entre les mots d'un même champs lexical Les mots d'un même lexique entretiennent entre eux des rapports sémantiques ou formels autres que la synonymie ou l'homonymie. Ce sont par exemple les liens :

- d'antonymie. Les antonymes sont en fait les 'contraires' des termes. Ils présentent

une symétrie de leurs caractéristiques sémantiques par rapport à un axe.
- d'hyperonymie/hyponymie. Les deux termes représentent un lien de généralisation/spécialisation. L'hyperonyme est le terme plus général d'un mot alors que l'hyponyme en est le terme plus spécialisé.
- de méronymie. La méronymie est une relation de partie à tout. Le méronyme désigne une sous-partie d'un mot.

Ces différentes relations ne seront pas étudiées de manière spécifique dans ce document.

2.4.2.3 Difficultés récurrentes des textes de biologie

Un des principaux problèmes qui fait de l'IEN en biologie une tâche complexe est l'absence de nomenclatures et de conventions d'écriture claires. Pour de très nombreux phénomènes en biologie, il n'existe aucun standard commun à l'appellation d'entités nouvellement élucidées et différentes communautés peuvent utiliser la même dénomination pour représenter deux entités distinctes. Plusieurs niveaux de complexité s'additionnent. D'une part, les conventions adoptées diffèrent entre les différents domaines d'expertise et les communautés qui composent la biologie. D'autre part, les nomenclatures établies pour certains types de concepts biomédicaux (par exemple, les allèles, les gènes et les protéines) sont différenciées selon l'organisme biologique étudié [ORS+02]. Il est à noter qu'il existe des tentatives d'harmonisation de la nomenclature dans des niches restreintes telles que les bases de données terminologiques **HUGO** et **Flybase**[12] qui définissent respectivement les standards des noms de gènes humains et de la drosophile. Néanmoins, ceci ne reste que des recommandations à l'intention des scientifiques et aucune obligation d'adoption de ces standards n'est à ce jour imposée. La validation du respect d'une nomenclature spécifique lors de la demande de parution d'un article s'effectue à la discrétion du journal scientifique concerné. Finalement, le nombre d'entités biologiques est très grand. Par exemple, et dans le cadre très restreint des gènes humains, **HUGO** recense au premier trimestre 2007 environ 24700 noms officiels de gènes, 29300 alias et 2970 dénominations obsolètes.

En théorie, les termes et les concepts manipulés devraient être mono-référentiels, chaque terme ne devant correspondance qu'à un seul concept et vice-versa. En pratique, les documents biomédicaux regorgent d'ambigüités.

[12] http ://flybase.bio.indiana.edu/

Evolution des nomenclatures De nouvelles entités biologiques sont constamment découvertes et nommées. A un instant donné le contenu des banques de données terminologiques ne peut correspondre aux dernières évolutions de la nomenclature en cours. De nombreux noms d'entités biologiques sont aussi supprimés de la nomenclature officielle lorsque ceux-ci sont jugés obsolètes. Ceci est particulièrement vrai lorsque la dénomination d'une entité repose sur des propriétés fonctionnelles qui par la suite ont été jugées erronées ou pas assez précises. Ces anciennes dénominations peuvent aussi être réutilisées pour nommer de nouvelles entités distinctes.

Noms complexes

- Nakagawa *et al.* [NM98] ont montré que plus de 85% des termes techniques spécifiques au domaine d'étude sont des mots composés mais non fusionnés. Les ENs peuvent être des bigrammes ou des trigrammes, néanmoins il n'est pas rare qu'elles soient plus complexes (par exemple, "Ras Guanine Nucleotide Exchange Factor SOS" et "Signal Tranducer and Activator of Transcription 3"). A l'opposé, de très nombreuses ENs utilisées en biologie moléculaire sont des mots uniques (par exemple, "actin" et "insulin") et parfois très courts (par exemple, "Vav" et "Nef").
- Il n'est pas rare non plus que ces ENs comportent des numéros ou des caractères non alphabétiques (par exemple, "p53", "Rho1p GDP/GTP exchange protein", "CD8+ Cell" ou "D1-cdk4"). Une étude [Kaz02] sur le corpus **GENIA** 1.0, spécifique à la biologie moléculaire, dévoile que la longueur moyenne des ENs du domaine est de 2,16 mots (avec un écart type de 1,40). La distribution observée est de 37% pour les mots uniques, 34% pour les bigrammes et 17% pour les trigrammes. Des cas d'ENs formées de 17 mots sont annotés dans le corpus (par exemple, "Double-positive (DP) CD (+) CD8 (+) CD3 (+/-) thymocytes").
- Dans certains cas précis des prépositions, conjonctions ou verbes peuvent faire partie intégrante de l'EN (par exemple, "Signal Tranducer and Activator of Transcription 3", "Rho1p GDP/GTP exchange protein for the Rho1p" et "Apoptotic Protease Activating Factor 3").

Il est a noter que les mots et *parts of speech* des ENs issus du domaine de la biologie moléculaire sont semble-t-il moins informatifs que ceux issus de domaines plus généraux. Nobata *et al.* [NCT00] ont montré que le pouvoir prédictif des *POS* au sein de modèles de REN est moindre dans les corpus spécialisés en biologie par rapport aux corpus généraux de type **MUC**[13].

[13] http ://www.muc.saic.com/

Est-ce vraiment un nom d'entité biologique ? La notion intuitive de ce qui constitue un nom d'entité devient confuse lorsque l'on observe les noms des entités biologiques présentes dans les textes, et plus particulièrement dans le domaine de la biologie moléculaire. Il est souvent ardu de différencier les noms d'entités biologiques de termes techniques ou de groupes nominaux réguliers (en anglais standard). Par exemple, chez *D. Melanogaster*, "yellow", "if" et "nervous" sont des noms de gènes. Plus l'expression employée pour se référer à une entité est fréquente et généralisée et plus celle-ci a des chances d'être nominalisée. Dans un domaine technique tel que la biologie moléculaire, il devient rapidement difficile de discerner les expressions nominalisées des mots environnants. De plus, il existe des situations dans lesquelles deux références 'étendues' à deux entités distinctes sont combinées afin de former une nouvelle référence à une troisième entité. Par exemple, le nom de la protéine "Ankyrin repeat and BTB domain containing protein 1" contient les références à deux autres entités distinctes, deux motifs protéiques. De même, le nom du gène "Testosterone-repressed prostate message-2" contient d'une part le nom d'une hormone et d'autre part le nom d'un organe interne. Une autre difficulté connexe est lié au phénomène dit de *noms en cascades* où l'association de *têtes* et/ou de *modificateurs* à un nom d'entité permet de représenter une deuxième entité, indépendante. Par exemple, "Cyclin" et "Cyclin dependent kinase" sont deux entités distinctes. De même, "Cyclin dependent kinase" et "Cyclin dependent kinase inhibitor" représentent deux entités différentes. De part cette nomenclature extrêmement permissive, les limites des noms des entités biologiques sont extrêmement floues, même pour un expert du domaine. Selon le contexte, l'expression "IFN regulating factor" peut désigner l'EN aussi appellée "IRF-1" ou décrire l'action de régulation de l'entité "IFN" sur un intermédiaire protéique.

Tuason *et al.* [TCL+04] ont calculé que la nomenclature des gènes de la souris et de la levure contienne environ 1,5% et 2,4% de termes de l'anglais standard, respectivement. Néanmoins ces chiffres relaivement bas sont sources de problèmes réels. Un nom de gène identique à un terme de l'anglais standard, apparaissant très souvent dans un texte et sans désambigüisation préalable, fait chuter sensiblement la précision des systèmes de REN et d'IEN.

Homonymie Un même nom peut se référer à différentes entités. Ce problème est fréquemment observé dans la dénominations d'objets biologiques non fonctionnellement liés entre différentes espèces animales ou entre différentes communautés scientifiques. Un autre exemple d'homonymie banale en biologie est le nom partagé à la fois par les gènes et les protéines. Il est parfois complexe de savoir si le nom est celui d'un gène ou d'un produit

du gène. Il est à noter que la tâche de désambigüisation des homonymes est ardue, même pour un lecteur humain. D'après une étude menée par Hatzivassiloglou *et al.* [HDR01], trois annotateurs spécialistes du domaine de la biologie moléculaire ne se sont accordés que 78% de fois sur la forme effective protéique, ARN ou gène des ENs à partir d'un corpus conséquent d'articles du domaine. Quelques conventions typographiques existent afin de distinguer certaines formes classiques d'homonymie et reposent sur la différenciation majuscule/minuscule et la mise en italique. Néanmoins peu de journaux en ligne respectent ces conventions typographiques et préfèrent réserver l'utilisation des italiques et des majuscules à d'autres fins, et notamment à la mise en relief du texte. De plus, les styles souligné, gras et italique sont très souvent perdus lors du formattage des publications dans les bases de données bibliographiques.

Tuason *et al.* [TCL+04] ont quantifié ces ambigüités pour les gènes au sein des ressources lexicales **Mouse Genome Informatics**, spécialisée dans la terminologie relative à la souris, **FlyBase**, pour la drosophile, **WormBase**, pour le vers, et **Saccharomyces Genome Database**, pour la levure. Les résultats montrent qu'entre 0 et 10% des noms au sein de chaque terminologie présentent des ambigüités. Le nombre d'ambiguïtés par nom allant de 2 à 10 et la plupart étant due à la présence de synonymes et autres alias et non à cause des noms officiels. Un facteur expliquant ceci est que certaines nomenclatures incluent intentionnellement des termes à la signification moins précise dans la liste des synonymes d'un gène. Un autre facteur est lié directement à la nomenclature établie pour chaque organisme. Ceci est uniquement vrai pour la souris et la drosophile. Les nomenclatures pour le vers et pour la levure sont beaucoup plus strictes que celles de la souris et de la drosophile. Il est à noter que des quatre organismes étudiés, la nomenclature des gènes humains se rapproche le plus de celle de la souris. Les seules restrictions imposées pour la souris sont l'utilisation d'un nom bref, précis et qui doit commencer par une lettre alphabétique. Les recommandations[14] en vigueur pour l'appelation de nouveaux gènes humains n'est pas beaucoup plus stringente. Les ambigüités de noms de gènes entre les différentes bases de données sont beaucoup plus nombreuses (entre 1 et 20% selon le couple de bases de données). Un autre facteur qui s'ajoutent à ceux précédemment cités ici est l'utilisation de noms identiques entre orthologues.

Hatzivassiloglou *et al.* [HDR01] proposent une méthode basée sur des techniques d'apprentissage automatique afin de tenter de résoudre un problème classique en biologie moléculaire : différencier la forme protéique, la forme ARN ou le gène d'une même entité.

[14] http ://www.genenames.org/guidelines.html

Certains éléments du contexte des occurrences connues des gènes, protéines et ARNs sont utilisé et statistiquement pondérés afin de discriminer les occurrences inconnues ultérieurement. Cette méthode est inspirée des travaux de Yarowsky [D.95] sur les textes d'anglais standard. Les mots présents dans l'environnement proche de l'entité à désambigüiser forment les *caractéristiques* qui servent de base à l'apprentissage. Cette information est raffinée en la complétant par l'observation de l'agencement des *caractéristiques* autour de l'entité (par exemple, "activates" devant une entité implique que l'entité est vraisemblablement un gène, à l'opposé, "activates" derrière l'entité suppose que celle-ci appartient plutôt à la classe des protéines) et notamment en prenant en compte la distance relative en mots des *caractéristiques* par rapport aux entités. D'autre part la morphologie des *caractéristiques* est aussi prise en compte. Ceci comprend la différenciation de la casse, les *parts of speech* associés et la racinisation. Les scores F obtenus par les auteurs varient entre 70 et 80% selon la classe a désambigüiser et la technique d'apprentissage utilisée. **AZuRE** [PCG+04] étend la problématique posée par Hatzivassiloglou *et al.* et permet de discerner les gènes des protéines mais aussi de distinguer les noms les gènes/protéines des objets non-gènes ou non-protéines.

Synonymie Il est courant qu'une molécule biologique soit nommée en fonction de sa fonction biologique particulière, à différents niveaux hiérarchiques (ex : une "ATP-dependent RNA helicase" est un sous type de "RNA helicase"), de sa similarité de séquence ou de la présence de motifs de séquence particuliers (ex : "DEAD/H Box-5"), sa masse moléculaire (ex : "protein p68"), ou encore de la combinaison de toutes ces situations (ex : "RNA helicase p68"). Néanmoins ces différentes nomenclatures ne sont pas exclusives et très généralement cohabitent afin de donner plusieurs noms synonymes à une molécule. Dans le cas des protéines, il est d'usage de leur donner le nom des gènes qui les codent. Les noms de protéines basés sur les gènes d'origine sont spécifiques à un organisme. En revanche, les noms de molécules contenant des références à des fonctions biologiques ou à des masses moléculaires sont généralement utilisés indépendamment de la taxonomie. Par exemple, "DRH1" et "DBP2" sont les noms d'une même protéine mais produite soit par *A. Thaliana* soit par *S. Cerevisiae* respectivement.

Acronymie Une variation de termes banale et partagée par l'ensemble des documents à caractère technique est l'utilisation d'acronymes. Il est toutefois important de signaler qu'il n'existe aucune règle stricte pour définir un acronyme en biologie à la différence d'autres domaines scientifiques. La définition d'un acronyme peut entrainer différents problèmes

précédemment relevés : tout d'abord la variabilité des formes synonymes d'acronymes. par exemple, "NFKB" et "NF Kappa B" sont deux formes acronymiques concurrentes de la protéine "Nuclear Factor-Kappa B" et retrouvées de manière indifférentes dans les textes. D'autre part la génération d'un acronyme peut créer une ambigüité homographique. Par exemple, l'acronyme "GR" sert à définir deux protéines distinctes, le "Glucocorticoid Receptor" et la "Glutathione Reductase".

ACROMED [PCC+01] est une des premières tentatives de résolution des acronymes dans le domaine des textes de biologie. Le système utilise une analyse syntaxique de surface afin de mettre en correspondance une forme développée et son acronyme. Yu *et al.* [YHRW02] ont mis au point un système adapté à la découverte et à l'identification des symboles et des noms complets des gènes ou des protéines à partir des textes. Ces travaux sont proches de ceux de Park *et al.* [PB01] et combinent à la fois des techniques de recherche de patrons et des méthodes à base de connaissances. Sur 30 résumés **Medline**, la précision et le rappel obtenus sont de 95% et 98%, respectivement. Pour un forme courte (abréviation, symbole ou acronyme) donnée, l'ensemble des formes longues associées candidates est extrait à la condition que les formes longues commencent par le même caractère que la forme courte. Chaque forme longue de cet ensemble est ensuite mis en correspondance avec la forme courte, en commençant par celle qui contient le moins de caractères, et subit 5 règles de recherche de patrons successives. Schwartz *et al.* [SH03] ont développé un algorithme simple, et plus généralisable que celui de Park *et al.*, dans le cadre de la désambigüisation des abréviations à partir des textes de biologie. Cet algorithme est communément utilisé au sein d'applications tierces et est adapté aux spécificités d'écriture des textes techniques en général. Les auteurs partent de l'hypothèse que la première occurrence d'une abréviation dans un texte est toujours accompagnée du nom développé, le premier apparait généralement entre parenthèses à la suite du second. La précision obtenue est de 95-96% et le rappel est de 82%. Narayanaswamy *et al.* [NRVS03] proposent une méthode extrêmement similaire à celle de Schwartz *et al.* qui permet de détecter sur un corpus de test différent 94,2% des paires. L'exactitude de l'identification est estimée à 87,4%.

2.4.2.4 Ressources terminologiques pour l'IEN

Plusieurs définitions et classifications pour les ressources terminologiques coexistent. Les ressources terminologiques peuvent être classées en trois grandes catégories : d'une part les vocabulaires contrôlés, d'autre part les ontologies spécifiques à une tâche unique, et

finalement les ontologies "pures". Les vocabulaires contrôlés sont indépendants de la tâche à accomplir et ne permettent pas de raisonnement logique. Le support du raisonnement logique donne la capacité à un concept donné d'être dérivé à partir d'un ou de plusieurs autres concepts. L'utilisation des ontologies spécialisées est limitée à des applications très spécifiques et supportent des formes de raisonnement logique. Les ontologies "pures" peuvent être ré-utilisées pour des tâches très hétérogènes et permettent encore une fois de raisonner logiquement.

Les vocabulaires contrôlés Les vocabulaires contrôlés définissent un ensemble de termes spécifiques au domaine et ont été développés afin de faciliter la diffusion des connaissances parmi les experts. Ces vocabulaires sont très couramment utilisés pour des tâches d'indexation bibliographique ou encore d'annotation de gènes. Ils sont généralement développés manuellement et structurés en arbres hiérarchiques où un nœud fils correspond à une sous-classe du nœud parent. Il existe un grand nombre de sources terminologiques de type vocabulaire contrôlé mises à disposition du public. La très grande majorité de ces sources sont en fait des bases de données intégratives et généralistes qui proposent de très nombreuses informations pour une famille particulière d'objets biologiques (par exemple, les séquences de gènes ou de protéines, leurs localisation chromosomique, des références à des orthologues, etc), dont des informations de nomenclature et de terminologie. A cette date, il n'existe pas encore de base de données terminologique exhaustive pour une classe ou un nombre restreint de classes d'objets biologiques. L'approche standard adoptée par les acteurs en IEN en biologie consiste à sélectionner un nombre variable de bases de données d'intérêt afin de constituer un dictionnaire local adapté à la tâche d'EI ou de RI envisagée. Les sources terminologiques les plus utilisées sont :

- **LocusLink**[15] [PM01] fournit une large quantité de noms officiels, de noms usuels, de symboles et d'alias ou synonymes pour un très grand nombre de gènes d'espèces animales variées. Depuis 2005 cette base de données a été intégrée dans **Entrez Gene** qui propose en outre certaines informations taxonomiques, structurelles, de localisation et de contexte génomique, bibliographiques et des pointeurs vers d'autres bases de données.
- **HUGO**[16] [WLD+04] est une base de données purement terminologique sur les gènes. De taille beaucoup plus restreinte que **LocusLink** par exemple, sa spécificité réside dans son contenu. Les noms et symboles retenus sont tous standardisés et officiels,

[15] http ://www.ncbi.nlm.nih.gov/entrez/query.fcgi ?DB=gene
[16] http ://www.gene.ucl.ac.uk/nomenclature/

l'organisme administrant **HUGO** ayant autorité sur la nomenclature des gènes humains. Elle propose aussi quelques pointeurs vers d'autres bases de données. **HUGO** est une initiative assez ancienne datant de la fin des années 70 mais qui a réellement pris son essor depuis quelques années uniquement. Il est intéressant de noter que d'après une étude de Tamames *et al.* [TV06], la nomenclature **HUGO** n'est que très peu respectée par les auteurs d'articles.

- **FlyBase**[17] [DC05] est la base de données de référence pour le génome de la drosophile. Les informations disponibles y sont nombreuses, très variées et recensent par exemple des informations sur les gènes et les allèles mutants, sur leur expression, sur les propriétés des transcrits et des protéines, sur les fonctions de leurs protéines, etc. Parmi ces différentes informations on peut retrouver les différents symboles, noms et alias ou synonymes d'un gène.
- **Mouse Genome Informatics**[18], **WormBase**[19] et **Saccharomyces Genome Database**[20] sont les pendants de **HUGO** et de **FlyBase** pour les organismes *M. Musculus*, *C. Elegans* et *S. Cerevisiae* respectivement.
- **Metathesaurus UMLS**[21] [Lin90] intègre une collection de plus de 60 vocabulaires à l'édition 2003. Les termes de ces vocabulaires sont groupés en deux niveaux. Tout d'abord, les termes provenant de vocabulaires différents et qui sont des variantes lexicales mutuelles. Par exemple, "interleukin 2", "Interleukin 2", "Interleukin 2 protein", "Il 2", "Il2" sont regroupés au sein d'un même concept. Ensuite, les concepts sont inter-connectés grâce à des relations qui décrivent des notions telles que "plus général", "similaire", "parent", "fils" ou "cousin" par exemple. Ces relations inter-concepts sont soient héritées des sources de vocabulaires d'origine ou générées par **UMLS**. Certaines sources de vocabulaire d'**UMLS** contextualisent les concepts qu'elles génèrent. Aussi **UMLS** ne réalise pas la fusion des diverses sources de vocabulaire au sein d'une hiérarchie unique et ainsi un même concept peut apparaître dans des hiérarchies d'**UMLS** différentes. Par exemple, le concept `fruit` appartient simultanément aux concepts `agriculture`, `food` et `plant component`.
- **GO (Gene Ontology)**[22] se concentre sur la mise au point de vocabulaires contrôlés inter-espèces : les fonctions moléculaires, les processus biologiques et les composants cellulaires. **GO** contient ainsi trois hiérarchies indépendantes, une pour chaque vo-

[17] http ://flybase.org/
[18] http ://www.informatics.jax.org/
[19] http ://www.flybase.org
[20] http ://www.yeastgenome.org/
[21] http ://www.nlm.nih.gov/research/umls/
[22] http ://www.geneontology.org/

cabulaire.

Les ontologies spécialisées La deuxième catégorie de ressources terminologiques, les ontologies spécifiques à une tâche, peuvent être encore sous-divisées en ontologies pour la description de schémas de bases de données et en ontologies pour la traduction de requêtes. Les ontologies utilisées au sein de **EcoCyc**[23] et **GDB**[24] sont des schémas de bases de données. Ces ontologies utilisent des méthodes orientées objet où chaque concept est défini en tant qu'instance de classe. Les relations exprimées sont définies explicitement d'une part entre les classes et leurs attributs et d'autre part entre paires de classes, rendant le schéma de la base de données à la fois plus flexible, plus extensible et plus lisible. **TAMBIS**[25], **BioKleisli** [DOTW97] et **BACIIS**[26] reposent sur des ontologies spécialisées pour réaliser des traductions de requêtes. Ces systèmes intègrent de nombreuses bases de données distribuées et hétérogènes. Les ontologies utilisées au sein de ces systèmes permettent de traduire des requêtes sur plusieurs bases de données en sous-requêtes spécifiques à une base de données particulière. Le processus repose sur deux étapes majeures. Les requêtes établies par les utilisateurs sont tout d'abord décomposées en sous-requêtes qui peuvent être traitées par une seule base de données. Ces sous-requêtes sont alors exprimées en utilisant des termes ontologiques. Les sous-requêtes sont finalement transformées en requêtes spécifiques à une base de données particulière, gérée par le système.

Les ontologies "pures" Elles sont réutilisables et indépendantes des applications qui l'uilisent. Encore très peu de systèmes terminologiques de ce type existent dans le domaine de la biologie. Il est a noter l'initiative récente **OBO**[27] dans ce domaine. **OBO** propose une collection d'ontologies spécialisées dont les méthodes de conceptions et les modèles sont partagés.

2.4.2.5 Ressources pour l'évaluation en REN

Deux types de ressources existent pour évaluer les performances des système de REN en biologie. D'une part, les jeux de données assemblés par les auteurs d'outils en REN eux

[23] http ://ecocyc.org/
[24] http ://www.gdb.org/
[25] http ://img.cs.man.ac.uk/tambis/index.html
[26] http ://mensa.sl.iupui.edu :8060/baciis/
[27] http ://obofoundry.org/

mêmes. Ils peuvent être alors mis à disposition du public conjointement avec la publication décrivant l'outil et les méthodes dont les performances ont été appréciées sur ce même ensemble de données annotées. D'autre part, des corpus conséquents, annotés et nettoyés par des experts du domaine, et dont la raison d'être fondamentale est l'utilisation lors de grandes compétitions internationales. De part leur grande taille et la qualité de leur annotation, ils sont devenus les étalons de référence *de facto* du domaine.

- Le corpus **GENIA**[28] [JDOTT03] est spécialisé dans le domaine de la biologie moléculaire. Il a été développé à partir de résumés disponibles dans la base de données **Medline** et extraits grâce aux termes **MeSH** "Human", "Blood Cells" et "Transcription Factors". Le corpus contient à ce jour environ 2000 textes annotés, contre 670 à sa création en 2000. Dans **GENIA**, seul un sous-ensemble des substances et de localisations biologiques impliquées dans des réactions biochimiques particulières avec les protéines est annoté. Cette sélection est basée sur une mini-ontologie propre à **GENIA** et correspond à 35 classes distinctes d'ENs. La compétition de REN **BioNLP 2004**[29] utilisait le corpus **GENIA**.
- Le corpus **BioCreative**[30] n'est annoté que pour une seule classe d'EN qui amalgame gènes et produits de gènes (protéines, ARN, ...). Il contient 75000 phrases d'entrainement et 2500 phrases d'évaluation.
- Le corpus **MedTag**[31] correspond à la fusion et à la mise à jour de trois différents corpus : **ABGene**, **MedPost** et **GENETAG**. Le résultat correspond à 25700 phrases, toutes issues de **Medline** et annotées pour une classe fusionnée unique gène/protéine.
- Le corpus **Yapex**[32] contient environ 200 résumés annotés provenant de **Medline**. La première moitié est réservée à l'apprentissage et la seconde moitié à l'évaluation. Seules les protéines sont annotées.
- Le corpus **LLL05**[33] est dédié à *Bacillus subtilis*. Il contient 2209 résumés où les gènes et protéines sont annotés mais non différenciés.

Il existe de nombreux autres corpus mis à disposition des acteurs de la FdT en biologie, néanmoins leur envergure ou leur utilisation reste limitée. Il est à noter que la très vaste majorité des corpus se concentre sur les protéines et les gènes uniquement.

[28] http ://www-tsujii.is.s.u-tokyo.ac.jp/~genia/topics/Corpus/genia-ontology.html
[29] http ://research.nii.ac.jp/~collier/workshops/JNLPBA04st.htm
[30] http ://www.pdg.cnb.uam.es/BioLINK/workshop_BioCreative_04/results/
[31] ftp ://ftp.ncbi.nlm.nih.gov/pub/lsmith/MedTag/
[32] http ://www.sics.se/humle/projects/prothalt/#data
[33] http ://genome.jouy.inra.fr/texte/LLLchallenge/#test_download

2.4.3 EI

L'extraction d'information, ou EI, est une tâche de FdT dont le but est d'extraire de l'information utile à partir d'un ensemble de textes. Une des applications de l'EI est d'aider les chercheurs qui doivent faire face à une quantité imposante de documents de recherche à analyser. L'EI peut être divisée en différentes sous-tâches telles que l'extraction de relations et l'extraction d'évènements séquentiels. Nous avons délibérément isolé la REN de l'EI, néanmoins, et classiquement, de nombreux auteurs tendent à considérer la REN comme une sous-tâche de l'EI. L'étape d'extraction de relations consiste en l'extraction de couples (ou tuples) d'ENs liées par une ou plusieurs relations cibles, par exemple une paire de protéines en interaction. L'extraction d'évènements séquentiels a pour but d'extraire des séquences de relations (qui dans ce cas représentent des évènements), par exemple des séquences d'interactions entre protéines.

2.4.3.1 Historique

Dès 2000, de nombreuses études se sont penchées sur le problème de la reconnaissance d'interactions entre les protéines et d'autres molécules. Les premiers travaux peuvent être grossièrement divisés en deux approches. Afin d'isoler des relations entre ENs, la première approche envisagée se base sur le principe de co-occurrence. La présence simultanée des deux entités et d'un mot convoyant la sémantique de l'association, dans une même portion de texte, est alors significative d'une relation. Cette approche produit néanmoins de très nombreuses relations erronées. Par exemple, le paire IL-2 et RhoA sont détectés comme étant associés dans les phrases "IL-2 activates IL2-Receptor and RhoA inhibits NFAT-dependent transcription" et "RhoA does not activate IL-2". La portée de ce type de stratégie reste très limitée. Les travaux de Stapley *et al.* [SB00] sont représentatifs de ce type d'approche. Une approche classique d'extraction d'information sur des réactions entre composés biologiques et chimiques est la recherche dans les phrases de patrons d'expressions régulières, façonnées manuellement, comportant un ensemble pré-défini de verbes pouvant être représentatifs de l'interaction [BAOV99, OHTT01, SPT98, NW99]. La principale difficulté rencontrée est la création d'autant de patrons que de formes textuelles représentant un évènement unique. L'extraction d'information à partir de patrons d'expressions régulières peut se révéler rapide et efficace sur un nombre restreints d'évènements et dans un domaine très précis, néanmoins la charge de travail nécessaire pour préparer les patrons devient rapidement insurmontable dans toute autre situation. Le

logiciel **PIES** [Won01] est un exemple caractéristique de ce type d'approche. Une des premières approches envisagées afin de simplifier la construction de patrons d'extraction a été d'extraire préalablement les groupes nominaux des phrases complexes. Les groupes nominaux étant relativement simples d'un point de vue de la structure, une analyse approfondie de la syntaxe n'est pas nécessaire pour les extraire. Cette première étape d'abstraction de la phrase permet de traiter à part les groupes nominaux et de limiter le nombre de patrons à créer. Des outils tels que **EDGAR** [RTWH00] ou **GeneScene** [LCM+03] utilisent le **Metathesaurus UMLS** afin d'identifier les entités candidates contenues dans les groupes nominaux extraits. Une fois ces groupes nominaux isolés de la phrase et identifiés, le processus d'extraction des relations peut se poursuivre. Parallèlement à ces méthodes centrées sur le design manuel de patrons d'extraction, de nombreuses techniques utilisant des méthodes d'apprentissage automatiques ont vu le jour. Celles-ci sont sensées soulager le travail manuel de construction de patrons. Une ensemble de données d'apprentissage est tout d'abord préparé. Il consiste majoritairement en un ensemble de phrases et de paires d'ENs unies par une ou des relations particulières, préalablement annotés et isolés manuellement par un expert du domaine. Cet ensemble de référence sert ensuite d'entrée au système d'apprentissage automatique et fournit en sortie des patrons d'extraction adaptés. Malheureusement, dans la plupart des cas, le coût de la conception manuelle des patrons est tout simplement transféré à la construction d'un ensemble de données d'apprentissage fiable et conséquent. Par exemple, les travaux de Thomas *et al.* [TMOP00] utilisent des Modèles de Markov Cachés afin d'acquérir les patrons d'extraction. Plus tard, des travaux se sont proposés de scinder la tâche de TALN à proprement parler de la tâche d'EI. La migration des systèmes ainsi développés à d'autres domaines devient plus simple. La décomposition de la structure, indépendante du domaine, de la phrase est elle réalisée exclusivement à l'étape de TALN, alors que l'étape d'extraction d'information peut être facilement adaptée aux problématiques propres au domaine du texte. Yakushiji *et al.* [YTMT01] ont été des précurseurs dans ce domaine pour les textes de biologie. Ici, la tâche de TALN est basée sur l'utilisation de grammaires généralistes. Temkin *et al.* [TG03] ont développé manuellement leur propre *grammaire hors contexte* à partir de 500 résumés **Medline** afin d'extraire des interactions entre protéines. L'analyse syntaxique grâce à ces grammaires permet de convertir la très grande variété de phrases qui décrivent un même évènement en *schémas de cas*, où l'évènement central est le verbe convoyant l'action avec pour arguments les sujets et objets. L'extraction de l'information est alors réalisée simplement en mettant en correspondance des patrons développés manuellement avec ces structures simplifiées, les *schémas de cas*. La variabilité syntaxique étant absorbée par les *schémas de cas*, la définition de patrons s'en trouve simplifiée

et accélérée. Certaines applications néanmoins proposent l'utilisation de *schémas de cas* sans utiliser d'analyse syntaxique au préalable. C'est le cas notamment de Blaschke *et al.* [BV02] et de leur outil **SUISEKI**. Ils utilisent directement des motifs de textes dans le but de construire leurs *schémas de cas*, sans étape intermédiaire, et ce dans le cadre restreint de la détection d'interactions entre protéines. Néanmoins, les auteurs constatent dans leur article que la majorité des problèmes rencontrés découle de difficultés d'ordre syntaxique. L'ensemble des méthodes décrites dans la littérature permet de situer la précision obtenue sur des publications biologiques entre 60 et 80%. Des approches hybrides sont apparues très récemment et permettent d'identifier des interactions entre entités à partir d'approches d'apprentissage automatique appliquées sur la structure des phrases. Bunescu *et al.* [BM05] ont étudié la question à partir d'une *grammaire de dépendance*. Afin d'extraire des relations entre entités les auteurs ont développé un noyau *SVM* qui utilise les caractéristiques du plus court chemin sur un arbre syntaxique en tant que paramètres de modèle. Erkan *et al.* [EOR07] reprennent les idées de Bunescu *et al.* et utilisent cette fois à la fois des modèles semi-supervisés et supervisés (*HMM* et *k plus proches voisins*) afin de retrouver des interactions entre protéines. L'avantage des méthodes semi-supervisées sur les méthodes purement supervisées est que les données d'acquisition peuvent provenir à la fois des textes annotés et non-annotés. Comme vu précédemment, l'annotation d'un ensemble de documents fiable et conséquent est une tâche fastidieuse. Pour la tâche d'extraction de relations non définies entre protéines, les auteurs obtiennent les meilleures performances à ce jour sur le corpus **AIMed**[34].

2.4.3.2 Difficultés récurrentes communes aux textes de différents domaines

L'incertitude et les hypothèses Il est très difficile de juger de la qualité de l'information extraite lorsque celle-ci rentre dans le registre du non-factuel. Dans le texte ceci peut transparaitre dans l'écriture des auteurs du document par le biais d'expressions (par exemple "it appears that the protein regulates the gene") ou de l'utilisation de termes *modulateurs* (par exemple "The protein may regulate the gene") exprimant l'incertitude. Dégager l'information factuelle des hypothèses est de première importance dans les textes techniques et notamment expérimentaux. Une question connexe est le caractère informatif des propositions négatives (par exemple, "the protein does not regulate the gene"). Dans tous les cas et une fois ces instances correctement détectées, le choix appartient à la personne collectant ces informations de les discriminer ou non.

[34] ftp ://ftp.cs.utexas.edu/pub/mooney/bio-data/

Les approximations numériques et scalaires relatifs La résolution des approximations numériques, par exemple "around 8 :00", est complexe et peut dépendre à la fois du thème du texte ainsi que de la subjectivité propre à chaque lecteur. Par exemple, "around 8 :00" peut être une plage horaire de durée extrêmement variable selon le lecteur, de même il n'existe pas de consensus pour déterminer les bornes limites afin d'interpréter l'information "a protein whose molecular weight is around 240 Kd". Le domaine spécifique du texte, et ses recommandations si elles existent, peuvent néanmoins soit *guider* le lecteur ou au contraire prohiber certaines valeurs. Les problèmes rencontrés sont identiques lorsque l'on souhaite résoudre les scalaires relatifs (par exemple "expensive, "heavy"). Les connaissances *a priori* du lecteur ou les règles du domaine du texte peuvent fournir une aide précieuse, par exemple "an expensive jet plane" coûte plus cher que "an expensive watch". L'utilisation d'une ontologie du domaine peut aider à résoudre ce type de difficultés.

Ellipses syntactiques et sémantiques Une ellipse syntactique est la non-représentation d'une information sémantique et qui est signalée par un "trou" syntactique. Par exemple, "In our last experiment, the antibody IgG bound the first epitope, but IgA did [bound the first epitope] too". L'ellipse sémantique est similaire mais sans la présence signal du "trou" syntactique associé. Par exemple, "The committee started with [a discussion of, a debate about] the drug approval issue". L'utilisation d'une ontologie peut aider à résoudre ces ellipses. Certains auteurs [NMB03] proposent des méthodes basées à la fois sur des ontologies et sur la détection d'entités textuelles spécifiques qui sont susceptibles d'être impliquées dans des ellipses. Par exemple, le verbe "start" déclenche très souvent des ellipses sémantiques, comme dans "start [eating]". Ce verbe est associé au sein d'un lexique à un ensemble d'évènements, l'évènement ellipse est résolu grâce au contexte, c'est à dire grâce à la collocation sémantique d'autres termes de la phrase.

2.4.3.3 Difficultés récurrentes des textes de biologie

A la différence des articles de la presse quotidienne (qui constitue la principale cible des applications de fouille de texte) la structure des phrases des articles de biologie, et biomédicaux en sens large, est globalement plus complexe. Certaines difficultés sont communes aux deux types de documents, d'autres sont spécifiques aux textes de biologie.

Variations des formes verbales La majorité des interactions d'intérêt entre entités biologiques implique la présence de verbes médiateurs, les verbes peuvent servir à qualifier la relation qui unie les sujets aux compléments d'objet ou d'autres arguments d'un verbe. Néanmoins, cette interaction n'est pas toujours représentée par la construction ordonnée sujet-verbe-complément. Le syntagme verbal peut subir de nombreuses transformations dont la passivisation (transformation à la voix passive, voir la figure 2.6) et la nominalisation (transformation d'un syntagme verbal en syntagme purement nominal, voir la figure 2.7). Par exemple, l'évènement qui décrit l'activation d'une protéine X par une protéine Y peut être représenté sous différentes formes, dont "X activates Y" (indicatif présent), "Y is activated by X" (forme passive), "X activating Y" (gérondif), "activation of Y by X" et "Y activation by X" (formes nominalisées).

Scénarios De surcroît, la structure de la phrase exprime une séquence ordonnée d'évènements. Par exemple, la phrase suivante "After RKIP is phosphorylated on serine-153 by a PKC-dependent mechanism, RKIP dissociates from its known target, RAF1, to associate with GRK2 and block its activity." représente une suite ordonnée d'évènements biologiques, décrivant 3 états et 1 réaction biologique :

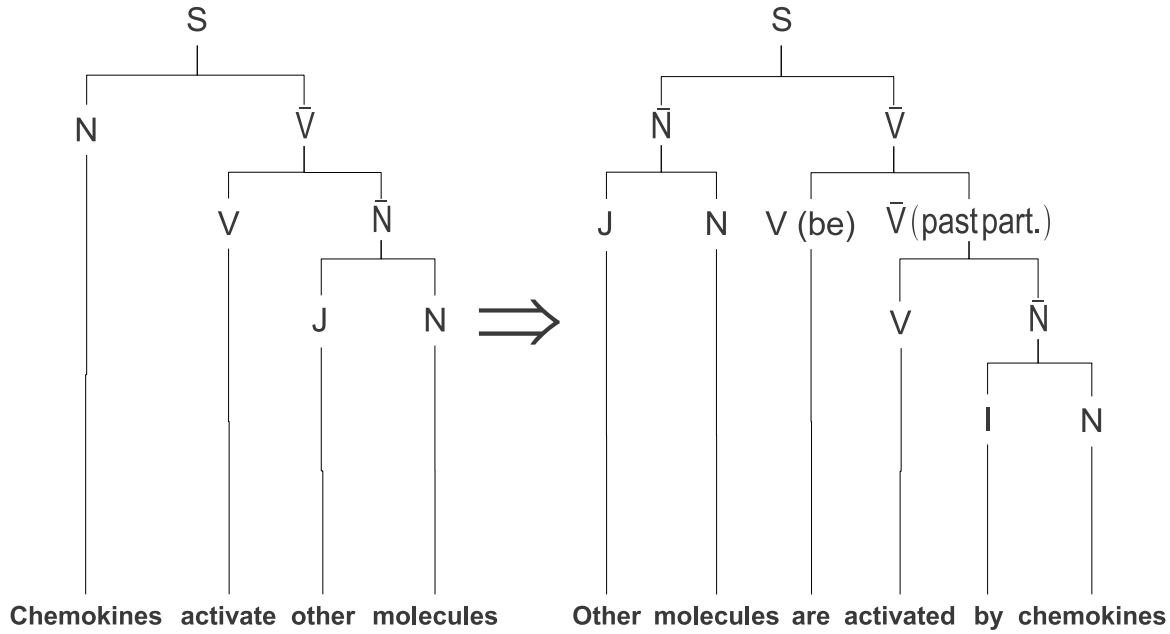
1. une sérine de RKIP, en position 153, est phosphorylée grâce à un mécanisme dépendant de PKC,
2. RKIP se dissocie de RAF1,
3. RKIP est associé à GRK2,
4. l'activité de RKIP est inhibée.

Ambiguïté des propositions relatives Ce problème n'est pas spécifique aux textes de biologie mais banal dans les publications scientifiques. Ainsi, dans la phrase "phorbol esters may induce posttranslational modifications of cellular transcription factors that alter their DNA-binding characteristics", la proposition introduite par "that" a-t-elle pour sujet "posttranslational modifications of cellular transcription factors" ou "cellular transcription factors" ?

Ambigüités des locutions prépositives Par exemple, dans la phrase "The experiment detects relevant tissues with antibodies", la locution "with antibodies" modifie-t-elle le groupe nominal "The experiment" ou "relevant tissues" ?

Ambiguïté des énumérations Les articles biomédicaux possèdent en règle générale des structures de coordination plus compliquées que celles présentes dans les textes journalistiques généraux [YTMT01]. Ce problème se rencontre très souvent dans les textes de biologie. Par exemple, dans la phrase "An experiment to detect antibodies in cells and tissues", "tissues" pourrait tout aussi bien être coordonné à "experiment", "antibodies" ou "tissues". Un autre problème connexe à l'utilisation des coordinations et très fréquemment rencontré en biologie est la ré-association ou non de parties de syntagmes aux membres coordonnés. Par exemple, dans la phrase "Lymphocytes produce either IL-4 or IL-2 Receptors"

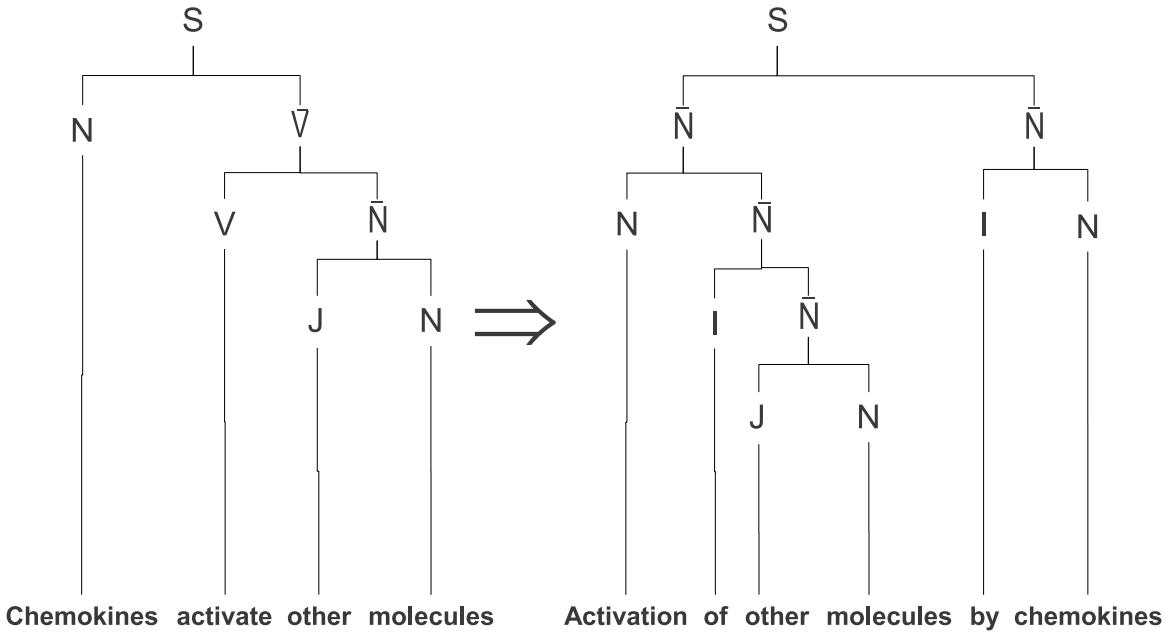
Fig. 2.6 – Exemple de passivisation en *grammaire de dépendance*



2.4.3.4 Analyse syntaxique superficielle ou complète ?

Afin d'identifier la nature des interactions entre entités biologiques dans le texte, les dépendances entre syntagmes d'une phrase (cf la sous-section 2.1.3) doivent être résolues. Deux approches sont alors classiquement opposées, d'une part l'analyse syntaxique dites superficielle, où l'on ne cherche qu'à identifier les constituants d'une phrase (groupes nominaux, verbes,...) sans spécifier leurs structures internes, ni leurs fonctions dans la phrase, et de l'autre l'analyse syntaxique complète ou profonde qui permet de générer un arbre syntaxique complet grâce à une grammaire de son choix. Chacune de ces deux

FIG. 2.7 – Exemple de nominalisation en *grammaire de dépendance*



approches possède ses propres avantages et inconvénients [Yak06]. Les méthodes dites profondes tendent à être plus coûteuses en temps de calcul et nécessitent d'avantage de mémoire. De plus il existe une polémique sur le fait que ce type d'approche produit plus d'ambigüités que les approches plus simples basées sur l'utilisation de patrons. L'analyse complète propose la structure complète d'une phrase alors qu'une analyse plus superficielle peut proposer une structure uniquement partielle de cette même phrase, en ignorant les portions de la phrase qui ne correspondent pas aux motifs des patrons. Finalement, la couverture des méthodes dites profondes reste inférieure à celle des méthodes dites superficielles à cause de leur complexité. Différentes tentatives d'utilisation d'analyse syntaxique superficielle dans le domaine de l'extraction d'information en biologie ont été réalisées [HDG00, TMOP00, PKK01]. Néanmoins, de l'aveu même des auteurs, les relations ainsi extraites restent simples. Yakushiji *et al.* [YTMT01] réalisent des pré-traitements de la phrase avant de la soumettre à une analyse syntaxique complète grâce à un analyseur syntaxique du domaine général. Ceux-ci permettent de résoudre certains problèmes d'ambigüités locales et d'augmenter l'efficacité de l'analyseur syntaxique. Le premier prétraitement réalisé a pour but de fusionner des groupes de mots d'un même syntagme nominal en une seule unité nominale que l'analyseur syntaxique pourra traiter correctement. Cette étape est très importante dans les textes techniques de biologie. Ils peuvent contenir des groupes nominaux très complexes et qui parfois défient la syntaxe anglaise

commune. Le deuxième pré-traitement lui est spécifique au problème de l'ambigüité lexicale des textes de spécialité. Le but ici est d'orienter l'assignation des *POS* à ceux plus fréquemment rencontrés dans les textes de spécialité étudiés par rapport aux textes généraux. Les entrées lexicales utilisées par l'analyseur syntaxique sont alors inscrites dans le cadre plus restreint défini par l'analyseur de surface employé par les auteurs. Les exemples suivants permettent de mieux distinguer les manques flagrants d'une analyse purement superficielle. Nous souhaitons extraire les paires d'entités ("A", "B"), ("C", "D") et ("E", "F") des phrases suivantes :

- "A activates B"
- "C is activated by D"
- "E is known to activate F"

Si les patrons d'extraction sont construits à partir des mots de surface, ceux ci peuvent se présenter sous les formes suivantes :

- "x activates y"
- "x is activated by y"
- "x is known to activate y"

Une forme généralisée de patron d'extraction peut être préparée à partir de ces trois formes différentes :

- "x␣(un ou plusieurs mots intermédiaires suivis d'un espace ou aucun mot intermédiaire)activate(un, plusieurs ou aucun caractère intermédiaire)␣(un ou plusieurs mots intermédiaires suivis d'un espace ou aucun mot intermédiaire)y"
  Ce motif est équivalent à l'expression régulière suivante :
  `{X}\ (([^\ ]+)\ )*activate[^\ ]*\ (([^\ ]+)\ )*{Y}`

Ce type de patron d'extraction peut détecter une relation qui n'existe pas à partir de la phrase "G activates H and I inhibits J". Connaître au préalable l'ensemble des mots qui peuvent être insérés entre les entités x et y et le mot "activate" afin de réduire le champs des relations possibles est, en soi, une tâche extrêmement ardue. Néanmoins, en procédant à l'analyse syntaxique des phrases d'exemple, nous pouvons obtenir les relations syntaxiques suivantes :

- "activate" a pour sujet "A" et pour objet "B"
- "activate" a pour sujet "D" et pour objet "C"
- "activate" a pour sujet "E" et pour objet "F"

En se basant uniquement sur ces relations syntaxiques, un patron d'extraction unique peut être construit :

- Si le verbe "activate" a pour sujet x et pour objet y alors (x, y) est une paire d'ENs unie par une relation d'intérêt.

Les avantages concrets sont de deux types : tout d'abord, les différentes représentations verbales de "activate" telles que "activates", "is activated by" et "is known to activate" ne donnent pas lieu à la création de patrons d'extraction différents. D'autre part, aucun sujet, objet ou verbe inopportun ne peut être inséré par erreur entre les entités x et y et le verbe "activate". Il est aussi intéressant de noter que les problèmes rencontrés par une analyse purement superficielle du texte proviennent du fait que les tâches spécifiques au TALN (par exemple, retrouver les différentes structures d'une phrase) et au domaine d'application (par exemple, trouver des mots significatifs d'une relation entre entités afin de générer des patrons d'extraction) sont entremêlées.

2.4.3.5 Représentation de l'information extraite

Ontologies La confrontation et la classification des données dans un système à base de connaissances est nécessaire pour répondre aux questions biologiques les plus complexes. Celles-ci peuvent typiquement nécessiter la mise en relation d'objets hétérogènes à partir de bases de données, tels que des relations entre séquences protéiques, de familles, structures et fonctions. L'utilisation d'une ontologie s'avère alors critique à l'inter-opérabilité et à l'intégration de telles données biologiques. Une ontologie est définie comme une collection de concepts représentant des entités spécifiques au domaine d'étude, l'ensemble des relations unissant les concepts et l'étendue des valeurs possibles pour chaque concept. Une ontologie peut alors être présentée grâce à différents formalismes, la représentation classique en EI étant le *schéma de cas*.

Schémas de cas Yakushiji *et al.* [YTMT01] utilisent des *schémas de cas* très simples avec un agentif, un datitif et un *cas* générique qui rassemble tout ce qui n'est pas agentif ou datitif. Leur outil repose préalablement sur l'analyse syntaxique complète de la phrase. Ils utilisent des règles de correspondance établies manuellement afin d'acquérir leurs *schémas de cas* à partir des structures syntaxiques générées par un analyseur syntaxique basé sur une *Grammaire syntagmatique dirigée par les têtes*. Néanmoins leurs règles demeurent assez simples et négligent les structures syntaxiques qui ne sont pas basées sur l'utilisation d'un verbe central. Néanmoins, l'efficacité de l'utilisation de pré-traitements spécifiques afin de contrebalancer certains défauts reprochés à l'analyse syntaxique complète a été démontrée dans leur article, à la fois d'un point de vue du temps de calcul, de l'occupation de la mémoire et de la couverture de texte analysé. Daraselia *et al.* [DYE+04] utilisent des triplets fixes de deux *schémas de cas* et d'un *cas* pour représenter l'information contenue

dans leur ontologie (voir la figure 2.8). Le *cas* défini une relation orientée entre les deux *schémas de cas*. Un *schéma de cas* se compose d'un nom unique et d'un ensemble de *cas*. Cette proposition de la structuration des relations entre entités biologiques leur permet de représenter un très nombre d'instances de l'information dans un environnement à la fois plus strict et normalisé. Ceci inclut l'activité enzymatique d'une protéine, la localisation cellulaire, les interactions protéine-protéine, l'organisation en complexes et la régulation de processus biologiques divers.

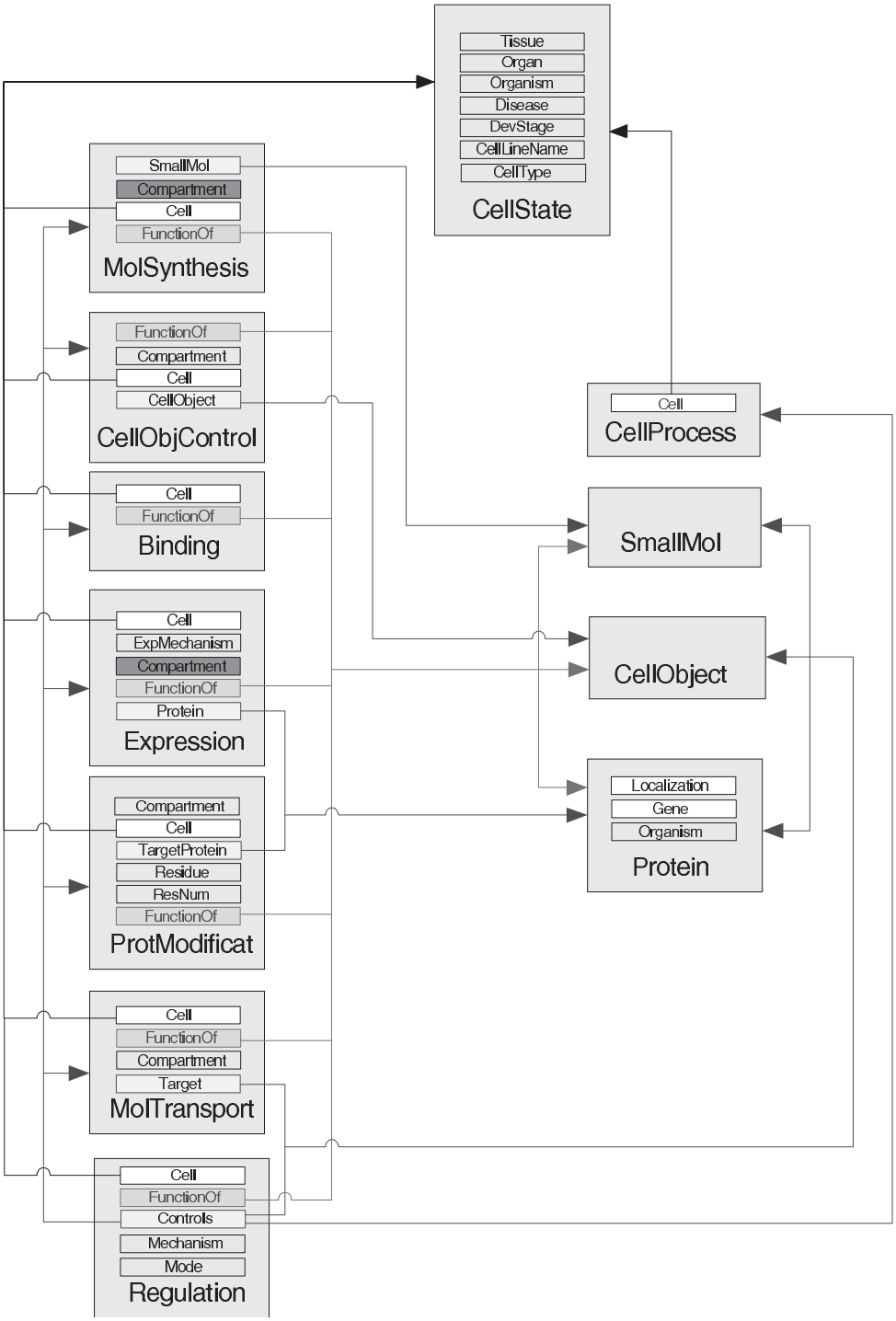
2.4.3.6 Ressources pour l'évaluation en EI

De la même manière que pour l'évaluation en REN, certains jeux de données sont très spécialisés alors que d'autres ont une vocation plus généraliste. Les premiers sont souvent mis au point afin de mesurer les performances et les spécificités d'un outil d'EI en particulier. Parmi les corpus de test de la deuxième catégorie, certains sont désormais considérés comme des jeux de données étalons dans le développement des outils d'EI génériques en biologie et sont pour la plupart régulièrement mis à jour.

- Le corpus **iProLink**[35][HMH+04], développé par le groupe **Protein Information Resource (PIR)**, se concentre exclusivement sur l'annotation des protéines et de leurs interactions. Des centaines de résumés et d'articles complets sont proposés. Chaque document est annoté avec des informations sur les modifications post-traductionnelles des protéines ainsi que des données en provenance de la base de données de séquences protéiques **PIR**.
- Le corpus **PASBio**[36][WSC04] propose un panel de relations annotées entre ENs très varié. En revanche, sa taille est assez réduite.
- Le corpus **AIMed**[37] contient environ 200 résumés **Medline** dont les instances d'interaction entre protéines ont été annotées.
- Le corpus **BioInfer**[38][PGH+07] comporte des interactions entre les protéines, les gènes et l'ARN. 1100 résumés ont été annotés. De plus, les types de relations détectées sont très variées. Ce corpus est parfaitement adapté à la mesure des performances d'outils généralistes dans le domaine de la biologie car le plus complet.

[35] http ://pir.georgetown.edu/iprolink
[36] http ://research.nii.ac.jp/ collier/projects/PASBio/
[37] ftp ://ftp.cs.utexas.edu/pub/mooney/bio-data/
[38] http ://www.it.utu.fi/BioInfer

Fig. 2.8 – Un survol de l'ontologie sur les fonctions protéiques utilisée par Daraselia *et al.* [DYE+04]



Les *schémas de cas* sont représentés par les grands rectangles. Les *cas* sont montrés en tant que rectangles de dimension plus réduite à l'intérieur des *schémas de cas* et sont connectés par des flèches aux *schémas de cas* cibles.

D'autres corpus existent, dont la taille et la spécialité varient. Ils restent cependant moins utilisés que les jeux de données majeurs présentés ci-dessus.

2.5 Résumé

Dans cette section, nous avons avons présenté un aperçu des applications de la FdT, et plus particulièrement de la tâche d'EI, pour la biologie. Nous avons montré les principales difficultés rencontrées dans les approches d'EI sur les textes de biologie et les solutions proposées dans la littérature du domaine afin de les résoudre.

Dans le suite du document nous présentons les méthodes et les résultats obtenus lors de cette thèse dans le cadre de l'extraction de connaissances sur les réseaux de la régulation des gènes.

Réalisation

Ce travail porte sur le développement de méthodes d'extraction d'information à partir des textes écrits en langage naturel dans le domaine de la biologie moléculaire. Plus précisément, le but de cette exploration du texte consiste à cartographier les réseaux de régulation transcriptionnelle de gènes ainsi que les mécanismes moléculaires sous-jacents chez l'homme et ce uniquement à partir des informations disséminées au sein des publications scientifiques.

La découverte et la mise en relation des réseaux de régulation de l'expression des gènes sont deux problématiques importantes de la biologie moderne. Comme nous l'avons vu dans la section 1.1, la compréhension des fonctions du monde du vivant passe notamment par l'élucidation des mécanismes moléculaires de la régulation transcriptionnelle des gènes et des interactions entre les différents acteurs de l'expression des gènes. Dans cette étude nous ne nous intéressons pas à la découverte de nouveaux systèmes de régulation de la transcription de gènes qui constitue le premier des deux axes de l'étude de la régulation des gènes. Parmi les techniques expérimentales récentes en biologie moléculaire, les *puces à ADN* et les protocoles de *ChIP-on-chip* ont offerts aux biologistes de nouvelles perspectives dans l'étude des mécanismes de la régulation de la transcription. Les gènes co-exprimés ainsi que les zones régulatrices de leurs promoteurs sont prédits à une échelle globale. Les données générées par ces méthodes expérimentales sont gigantesques rendant leur analyse ardue. Nous travaillons ici sur l'agrégation et la mise en relation des données existantes du domaine. La littérature scientifique du domaine de la biologie moléculaire est une ressource extrêmement riche pour l'étude de l'expression des gènes. L'ensemble des connaissances du domaine y est accumulé [SF03] (voir le chapitre 2). Ces données sont présentées sous la forme de résultats expérimentaux (notamment ceux issus des expériences de *puces à ADN* et de *ChIP-on-chip*) ou de vues intégratives et critiques que ce soit sur les dernières avancées du domaine ou sur les connaissances de base de la spécialité.

Néanmoins, de part la taille conséquente de la littérature et de sa constante croissance, des méthodes automatiques doivent être mises au point afin d'explorer systématiquement ces données. Un des défis importants de la biologie moderne est de réussir à mettre en perspective les nouvelles découvertes avec ce qui est considéré comme déjà connu. Ainsi, un effort particulier doit être produit afin de développer des méthodes automatiques pour assister les biologistes à agréger et à analyser les données du domaine. Nous avons développé une application de FdT capable d'extraire et de collecter des données d'expression de gènes détaillées à partir des textes écrits en anglais. Nous présentons dans ce chapitre les différentes méthodes mis en œuvre afin de répondre à cette problématique ainsi que leur apport à la discipline de la bioinformatique.

Les modèles de connaissance de l'information à extraire

Cette présente section détaille les différentes données relatives à la régulation de la transcription de gènes que nous cherchons à extraire des textes. Comme précédemment exposé dans la section 1.2, le mécanisme de l'expression des gènes est un processus biologique complexe difficile à modéliser. Il implique de différents niveaux d'interactions entre les protéines, d'interactions entre l'ADN et les protéines et d'interactions génétiques. Les facteurs de transcription reconnaissent spécifiquement des séquences nucléiques particulières des zones promotrices des gènes. L'interaction efficace des facteurs de transcription avec ces séquences peut soit conduire à l'expression du gène concerné, c'est à dire à la production d'ARNm reflétant l'information portée par le gène, ou à sa répression, c'est à dire à l'arrêt ou au blocage de la production d'ARNm. L'issue de cette interaction dépend principalement de la nature et de la combinaison des entités biologiques impliquées. En retour, la transcription des gènes qui codent des facteurs de transcription est aussi sous l'influence de facteurs de transcription.

Dans cette étude, nous nous intéressons à deux questions biologiques complémentaires :

- d'une part, nous nous efforçons de mettre au point les méthodes de FdT nécessaires à l'établissement de cartes de réseaux de régulation de la transcription des gènes. En d'autres termes, nous cherchons à dresser la liste des gènes qui régulent positivement ou négativement la transcription d'un gène d'intérêt.
- D'autre part, nous nous concentrons sur le problème du décryptage des différentes étapes moléculaires qui conduisent à l'action effectrice des facteurs de transcription activés sur la régulation de l'expression transcriptionnelle d'un gène d'intérêt.

Un modèle simplifié de l'expression de gènes

Nous avons intégré les deux facettes de l'exploration de la régulation de la transcription des gènes au sein d'un même modèle simplifié. Celui ci est présenté dans la figure 2.9.

Notre modèle est centré sur la régulation spécifique d'un gène donné par un autre (ou par lui même) au sein d'un type ou d'une lignée cellulaire particulière. Il existe donc autant d'instances du modèle que de combinaisons entre gène régulateur, gène régulé et le type cellulaire siège de la régulation.

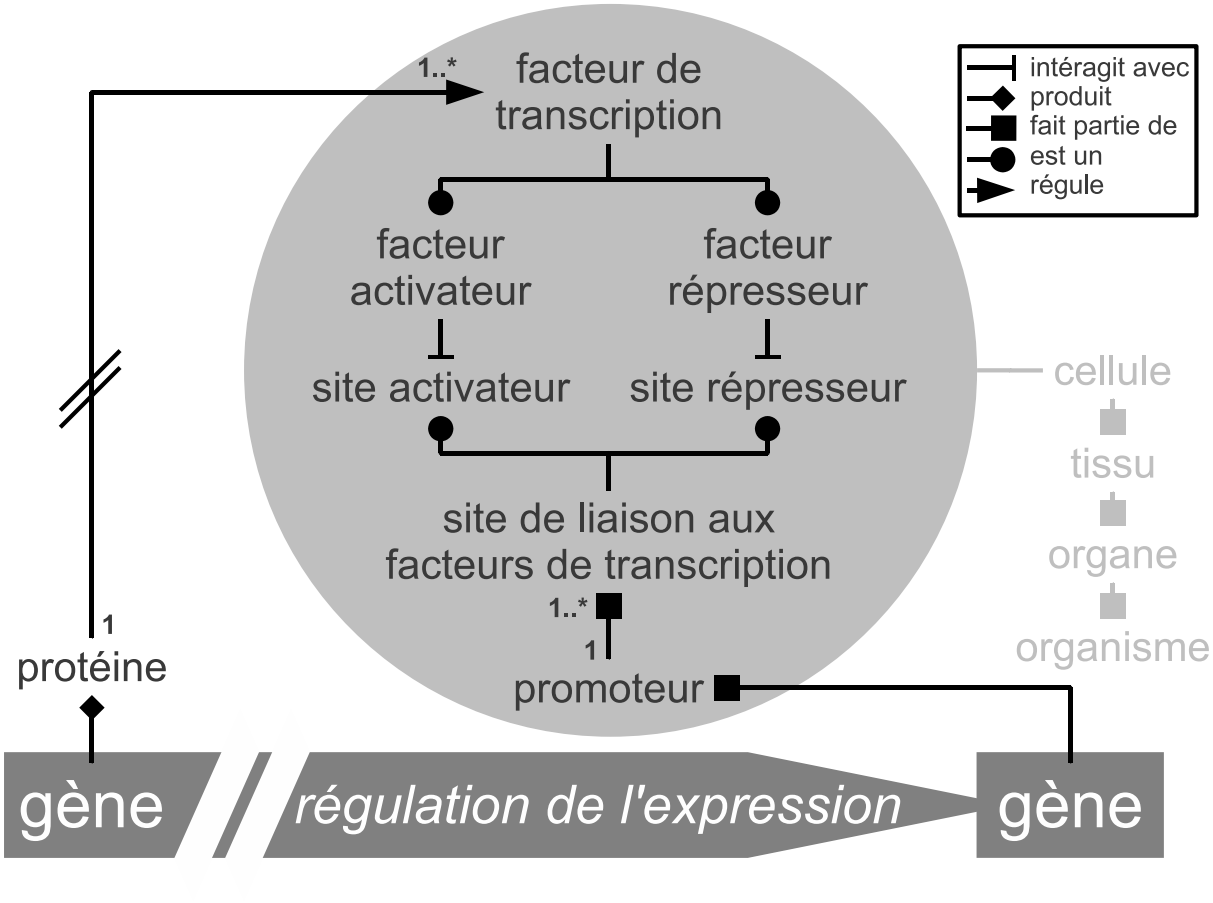


FIG. 2.9 – Diagramme simplifié du mécanisme de la régulation de l'expression de gènes : les entités biologiques impliquées et leurs relations

Au sein de notre modèle, nous résumons l'action du gène régulateur sur le gène régulé par ce que nous considérons être l'un des aspects fondamentaux de la transcription eucaryote : la combinatoire des interactions entre les facteurs de transcription mobilisés lors de l'établissement du signal de régulation par le gène régulateur et les sites de liaison aux facteurs de transcription du promoteur du gène régulé.

La voie de régulation de la transcription ainsi modélisée permet de prendre en compte :

1. le catalogue des facteurs de transcription mobilisés lors la régulation du gène,
2. l'interaction spécifique d'un facteur de transcription référencé sur tel ou tel site de liaison aux facteurs de transcription

Bien qu'une instance du modèle soit définie par le trio gène régulé, gène régulateur et cellule cible, certains paramètres utilisés dans le modèle ne dépendent pas simultanément de ces trois contextes. Certaines entités et leur propriétés ainsi que certaines relations peuvent être partagées entre plusieurs instances du modèle. Par exemple, l'action soit activatrice ou répressive sur la transcription d'un site de liaison aux facteurs de transcription est globalement conditionnée au gène régulé et non au gène régulateur ou au type cellulaire.

Dans ce modèle simplifié, nous passons notamment sous silence les phénomènes épigénétiques tels que la méthylation de l'ADN. En ce sens, c'est la raison pour laquelle nous considérons chaque voie de transduction du signal de régulation de la transcription dans son propre contexte cellulaire. Il est aussi à noter que toutes les données cinétiques ont été elles aussi omises, le but en est encore une fois de simplifier la modélisation du processus. Ainsi, par exemple, nous ne pouvons distinguer les différences d'expression d'un même gène au cours de la vie de la cellule.

Deux difficultés majeures doivent être prises en considération lors de la découverte des paramètres du modèle à partir des textes :

- tout d'abord, nous devons utiliser les données parcellaires ou incomplètes présentes dans les textes. Celles-ci peuvent être classées en trois catégories :
  - d'une part, le contexte d'une relation ou de l'implication d'une entité particulière peut être incomplet voire absent. L'indication du gène régulé, du gène régulateur et du type cellulaire nécessaires à la définition de ce paramètre du modèle peut ne pas être précisé dans les textes ou ne pas être détecté. Par exemple, un facteur de transcription spécifique aide à l'activation de la transcription d'un gène d'intérêt grâce à son interaction avec un site de liaison donné, or le contexte cellulaire de l'interaction n'a pu être élucidé (dans quelle cellule ?). Dans ce document nous passerons sous silence cette difficulté particulière et nous ne chercherons pas à confronter ces données fragmentaires, qu'elles soient issues d'un même article scientifique ou en provenance de différentes publications.
  - D'autre part, il est possible que la relation qui anime deux entités soit imprécise. La relation sous-jacente est ici compatible avec le modèle mais n'est pas explicite.

Par exemple, dans un type cellulaire particulier, un facteur de transcription donné interagit spécifiquement avec un site de liaison d'un gène d'intérêt, néanmoins la conséquence de cette interaction reste incertaine (activation ou répression de la transcription ?).

- Finalement, certaines ellipses et raccourcis dans l'établissement du modèle sont tolérés. Le chemin exact entre deux acteurs du modèle est ici inconnu, les noeuds intermédiaires ne sont pas précisés. Par exemple, nous savons qu'une protéine donnée va être à l'origine du signal d'activation de la transcription d'un gène d'intérêt dans une cellule particulière. En revanche nous ne connaissons pas les facteurs de transcription mobilisés lors du processus, ni les sites de liaison aux facteurs de transcription impliqués.

– Ensuite, nous devons gérer la présence de données contradictoires ou conflictuelles. Certaines règles de décision sont nécessaires afin de retenir le paramètre correct du modèle lorsqu'il est décrit par des données opposées. Ce type de situation peut être retrouvé lorsque l'on confronte les informations proposées par différents auteurs et articles. La situation peut aussi survenir au sein d'un même document, ici ces données sont décrites dans des contextes différents, parfois extrêmement fins et complexes. Par exemple, lors de la lecture de la première section d'un article scientifique, nous apprenons qu'une protéine particulière permet l'activation d'un gène spécifique, or cette même information est contredite dans la deuxième portion du document.

Les moyens mis en œuvre afin d'extraire l'information relative à la régulation de gènes dans les textes ainsi que sa structuration sont développés dans les sections suivantes.

Le processus de FdT

Nous avons développé un outil complet d'EI afin de prendre en charge l'ensemble de la chaine de traitement de l'information relative à la régulation de gènes à partir des documents scientifiques.

Les documents textuels sources peuvent être aussi bien des articles non structurés (texte brut) que des articles semi-structurés (XML **Medline**[39] et **BioMed Central**[40]). Comme précédemment évoqué dans la littérature [SWS+04], les différentes sections d'un

[39] http ://www.nlm.nih.gov/databases/dtd/nlmmedline_021101.dtd
[40] http ://www.biomedcentral.com/xml/article.dtd

article scientifique en biologie ne contiennent pas les mêmes informations et les mêmes catégories de données. De même, la densité d'information varie d'une section à une autre. En revanche, nous ne nous limitons pas aux seuls résumés. L'étude précédente montre que le résumé contient deux fois moins de concepts que dans le reste du document. Or l'information que nous cherchons à extraire est relativement fine et non forcément présentée dans les résumés. Nous travaillons sur les textes complets mais nous choisissons de ne pas traiter les parties expérimentales dédiées aux matériels et méthodes si elles sont présentes. Deux raisons majeures sont à l'origine de l'exclusion de ces rubriques : d'une part elles sont les plus pauvres en information [SWS+04] et d'autre part leur contenu est le plus technique et donc le plus difficile à acquérir car la syntaxe est hautement spécialisée (voir la section 4.1).

Le processus d'extraction de l'information pertinente à partir de ces textes est alors décomposé en deux grandes étapes, tout d'abord la REN, décrite dans la section **??**, puis l'extraction des relations d'intérêt entre ENs, décrite dans la section 4. La première étape dite de REN (voir la section 2.4.2) nous permet d'identifier l'ensemble des acteurs du modèle de la régulation de la transcription des gènes manipulés dans les textes. La deuxième étape est nécessaire à la mise en relation des différents acteurs du modèle détectés lors de la REN. Seuls les liens fonctionnels tissés entre les ENs identifiées et compatibles avec les paramètres imposés par le modèle de régulation de la transcription des gènes sont alors extraits.

L'ensemble des informations modélisées est alors structuré à l'aide d'XML ou répertoriée dans une base de données. Les données ainsi rassemblées permettent de définir, en un sens, une ontologie simplifiée du domaine.

Dans le premier chapitre, nous proposons une approche de REN basée sur l'utilisation de dictionnaires et appliquée à la découverte des ENs manipulées dans notre modèle de la régulation des gènes. Nous décrirons les étapes de création des dictionnaires adaptés à cette tâche ainsi que les solutions apportées aux problèmes de la variabilité des noms des entités biologiques qui peuvent être mis en œuvre à cette étape. Nous proposerons ensuite une méthode d'EEN basée sur une grammaire spécialisée suivi des techniques mises en place pour l'IEN à partir des ENs extraites et utilisant les dictionnaires développés. Une section particulière sera consacrée aux difficultés liées à l'ambigüité des ENs extraites des textes et des techniques proposées afin d'y répondre.

Dans le deuxième chapitre, nous présentons les méthodes utilisées afin de détecter

et d'identifier les relations d'intérêt dans le cadre de notre modèle de la régulation des gènes entre les ENs des textes. Dans un premier temps nous décrirons les méthodes liées à l'acquisition de la syntaxe à partir des textes et des problèmes liés à la diversité des représentations syntaxiques pour exprimer une même information. Nous proposerons une approche basée sur l'utilisation de *structures prédicat-arguments* dans le but de simplifier la syntaxe et de nous affranchir de ces difficultés dans la mesure du possible. Dans une deuxième partie, nous détaillerons les méthodes nécessaires pour découvrir les relations fonctionnelles qui unissent les objets biologiques d'intérêt à partir des *structures prédicat-arguments* grâce à l'injection de connaissances terminologiques propres au domaine de l'étude. Nous décrirons les différentes étapes successives dites de conceptualisation qui nous permettrons d'identifier les interactions entre objets biologiques présents notre modèle de la régulation de gènes.

Travaux apparentés

Concernant les travaux d'EI dans le domaine de la régulation de gènes, seules deux études sont apparentées aux nôtres.

- La première [SJO+06] a pour but l'extraction de réseaux de régulation de gènes à partir des textes. Les auteurs montrent qu'une grande exactitude lors de l'EI peut être atteinte une fois les barrières terminologiques spécifiques au domaine d'étude dépassées. Ils ont ainsi été capables de régler finement le problème classique de la découverte d'interactions entre protéines aux spécificités de l'expression de gènes. Une différence essentielle avec notre approche consiste en la relative généricité de leur système vis-à-vis du modèle animal utilisé dans les textes. L'ensemble de nos méthodes et le contenu de nos dictionnaires sont centrés sur l'espèce humaine (ou à défaut des mammifères), en comparaison leur étude porte simultanément sur la levure, la souris et la bactérie *B. Subtilis*. Il est à noter qu'en aucun cas les techniques mises au point par les auteurs ne permettent de différencier les données extraites en fonction de l'espèce concernée, il est ici uniquement question d'étendre globalement la couverture de l'étude aux subtilités et aux spécificités d'autres modèles animaux. Néanmoins, leur étude se limite à déterminer quelles protéines régulent la transcription de gènes. Notre modèle de la transcription de gènes est plus complexe et plus nuancé. Nous proposons ici d'aller un peu plus loin et de fournir en supplément des détails (notamment des donnes expérimentales) sur le processus effectif de la transcription de gènes ainsi que son contexte cellulaire. Leur système est basé sur

une approche à base de règles, définies par des experts, afin de capturer les relations sémantiques entre les gènes et les protéines dans le cadre la régulation de la transcription. L'étape de REN est quant à elle assurée par l'utilisation de techniques similaires aux nôtres qui nécessitent notamment le recours à des dictionnaires.

– L'autre étude [PZC+04] se focalise sur la découverte des associations entre facteurs de transcription, catégories **GO** et situations pathologiques dans les textes. Bien que ce soit un des rares systèmes à s'intéresser à des problèmes liés à l'expression de gènes, le but des auteurs est très éloigné du nôtre. L'application qu'ils ont développée fournit un aperçu crédible des relations fonctionnelles qui existent entre ces trois catégories d'acteurs, tandis que nous nous concentrons plutôt sur l'aspect moléculaire des interactions.

Chapitre 3

REN

Dans ce chapitre, nous présentons une approche globale de REN proposant une méthode de EEN couplée à une technique d'IEN basée sur l'utilisation de dictionnaires. Dans un premier temps, nous détaillerons le processus de création de dictionnaires utilisés afin d'identifier les objets biologiques manipulées dans notre modèle de la régulation de gènes puis, dans un deuxième temps, nous décrirons les méthodes d'EEN et d'IEN mises en œuvre afin de découvrir les ENs des dictionnaires à partir des publications scientifiques. La mesure des performances du système de REN sur un jeu de données de biologie moléculaire nous permettra de conclure cette partie.

3.1 Dictionnaires d'ENs

Nous proposons dans cette section une méthode de construction de dictionnaires contrôlés pour l'IEN. Nous nous limitons ici volontairement à définir des dictionnaires servant à identifier :

- le nom des gènes,
- des protéines,
- des souches ou des lignées cellulaires,
- des tissus ou des organes,
- des expériences ou protocoles expérimentaux,
- des sites de liaison aux facteurs de transcription,
- et des facteurs de transcription

dans les publications scientifiques traitant de la biologie humaine. Nous allons décrire d'une part les diverses sources terminologiques nécessaires à leur création et d'autre part nous nous efforcerons d'expliquer les différentes techniques mises en place afin de répondre aux difficultés spécifiques de la nomenclature du domaine. Le processus de création des dictionnaires est schématiquement proposé dans la figure 3.1. La couverture relative de nos dictionnaires, après expertise du contenu, est évaluée à partir d'un sous-ensemble du corpus de référence **GENIA**.

Afin de détecter et de reconnaitre les termes correspondant à des concepts biologiques nous avons opté pour une approche à base de dictionnaires. Cette technique nous permet de mettre aisément en correspondance des ENs détectées dans les textes et les objets biologiques référencés dans les banques de données biologiques.

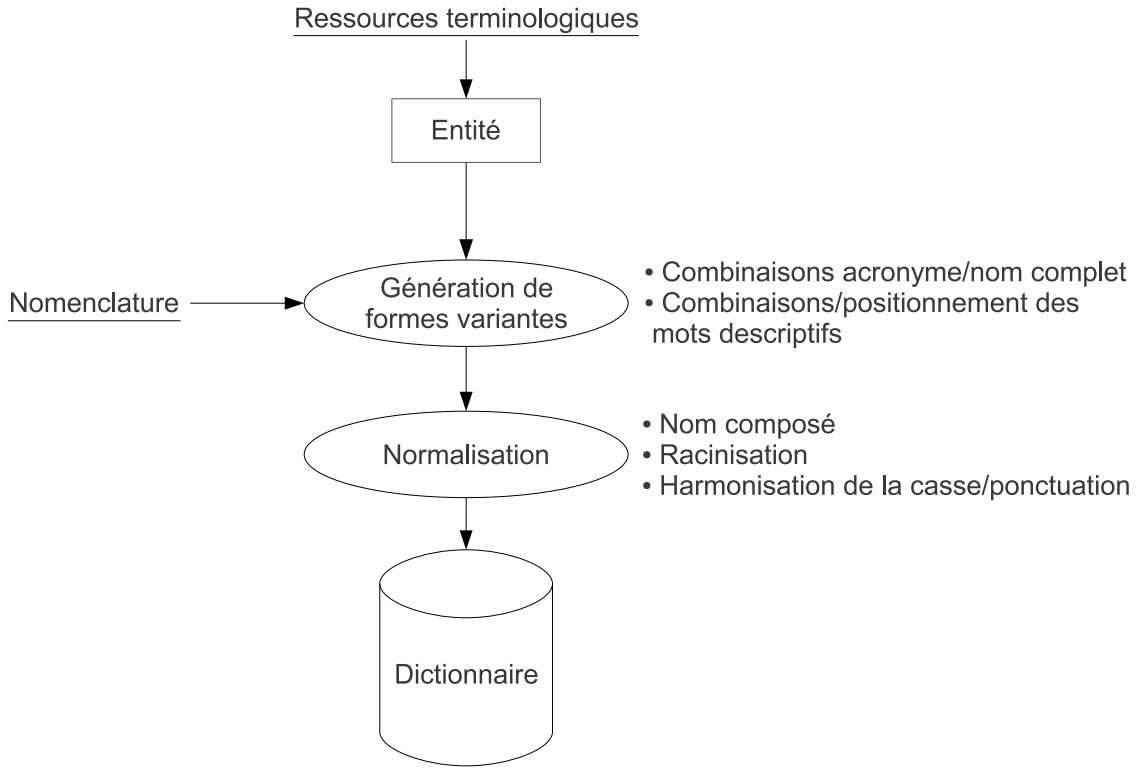


FIG. 3.1 – Processus de création des dictionnaires

3.1.1 Ressources utilisées

Nous avons sélectionné un nombre restreint mais complet et correctement expertisé de bases de données terminologiques publiques afin de construire nos différents dictionnaires :

- **LocusLink**, **HUGO** et **OMIM** ont servi à référencer les gènes et protéines,
- **TRRDSITE**[1] les sites de liaison aux facteurs de transcription,
- **TFD**[2], **COMPEL**[3], **TRRDFACTORS**[4] et **TFFACTORS**[4] les facteurs de transcription,
- **Metathesaurus UMLS** les cellules, tissus, organes, protocoles expérimentaux et techniques et appareillages.

Ces banques de données terminologiques ont été sélectionnées d'une part parce qu'elles répondent à nos besoins et d'autre part parce que leur contenu présente une qualité d'expertise certaine. Les critères retenus pour leur sélection sont :

- la quantité d'information qu'elles contiennent. Seules les banques majeures pour une famille particulière d'objet biologique ont été retenues. Plusieurs banques de données peuvent être fusionnées afin de potentiellement augmenter le nombre de formes synonymes disponibles dans nos dictionnaires si la troisième condition exposée ci-dessous est remplie.
- La fréquence de mise à jour. Les banques de données qui n'ont pas été mises à jour plusieurs mois auparavant n'ont pas été retenues. Comme précédemment exposé dans la section 2.4.2.3, la prise en charge de l'évolution des nomenclatures est un facteur important dans la qualité des dictionnaires générés.
- La présence de liens d'identité entre les banques de données pour les objets manipulés. S'il existe plus d'une banque de données terminologique pour une famille d'entité biologique spécifique, il est important que les différentes entités identiques entre les banques de données soient identifiées. Cela évite la présence de doublons dans les dictionnaires résultants. Comme expliqué un peu plus loin, les entités sont classées en fonction de leur identifiant. Si l'identifiant est différent entre deux banques terminologiques mais représentent la même entité biologique, l'absence de marqueur d'identité donne lieu à l'enregistrement dans les dictionnaires de deux entités indépendantes qui, si elles partagent des dénominations en commun, seront étiquetées en tant qu'homonymes et non en tant que synonymes.

Ces bases de données proposent pour chaque entrée à la fois un ou plusieurs noms complets usuels ainsi qu'un ou plusieurs symboles/acronymes/abréviations officiels ou non définissant un même objet biologique. Ces différents alias peuvent être encore en usage ou ne plus avoir cours dans la littérature contemporaine. Les différentes entrées sont fusionnées grâce aux symboles **HUGO** communs pour les gènes et les protéines. Dans le

[1] http ://wwwmgs.bionet.nsc.ru/mgs/gnw/trrd/
[2] gopher ://gopher.nih.gov/77/gopherlib/indices/tfd/index/
[3] http ://compel.bionet.nsc.ru/new/index.html
[4] http ://www.gene-regulation.com/pub/databases.html#transfac

cas des facteurs de transcription, aucune banque de données seule ne s'est révélée assez complète pour répondre à nos besoins. Une inspection manuelle des différentes ressources relatives aux facteurs de transcription sélectionnées a montrée que leur contenu, comparé deux à deux, est exclusif dans une proportion de 20 à 50%. Or il n'existe aucun marqueur relationnel entre les différentes banques de données pour les facteurs de transcription. Nous avons néanmoins décidé de fusionner ces ressources grâce aux noms complets des entités manipulées. Les facteurs de transcription formant une famille particulière de protéines, nous apparions les entrées relatives aux protéines et aux facteurs de transcription lorsqu'au moins au alias (nom complet) est commun. Les alias ainsi mis en commun ne sont plus ceux d'une protéine mais ceux d'un facteur de transcription.

Autant que possible, les entrées conflictuelles ou orphelines sont examinées manuellement. Nous ne nous intéressons qu'aux entités d'origine humaine ou à défaut de mammifère. Nous utilisons pour cela les informations contenues dans les champs relatifs à l'espèce ou à la classe dans les bases de données multi-espèces. Contrairement à d'autres approches similaires [KT04], les différents objets manipulés dans les dictionnaires n'entretiennent entre eux aucune relation d'appartenance ni de composition. L'information n'est présente et utilisable automatiquement qu'à partir des bases de données **TRRD-FACTORS** et **TRRSITE**. Nous avons préféré ignorer ces données dans un souci de généricité des méthodes de construction des dictionnaires.

3.1.2 Composition et description

3.1.2.1 Les variantes de noms

La capacité d'un système de REN à reconnaitre les formes variantes d'une EN (voir le paragraphe 2.4.2.2) est déterminante dans la couverture des textes et en conséquence apparait comme un des facteurs les plus importants lors de l'optimisation du score de rappel. Nous stockons dans nos dictionnaires les différentes représentations lexico-sémantiques, grammaticales et syntaxiques de chaque objet biologique d'intérêt à partir des différentes bases de données sources. Pour l'ensemble des bases de données, à l'exception du **Metathesaurus**, chaque tuple peut contenir plusieurs alias d'une même entité. Chacune des banques de données utilisée possède son propre format d'entrées. Par exemple, **HUGO** préconise aux contributeurs de sa base de données de suivre les recommandations d'écriture suivantes :

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

Chaque alias est alors considéré comme un forme variante valide d'une même entité et constitue une entrée à part entière lors de la création de nos dictionnaires.

3.1.2.2 Génération de formes variantes

Le panel des formes variantes répertoriées dans nos dictionnaires et issues des bases de données terminologiques est insuffisant pour couvrir l'ensemble des variantes utilisées par les auteurs d'articles scientifiques. Il n'est en effet pas rare que de très subtiles modulations (par exemple, l'utilisation de formes acronymiques mixtes ou de l'inversion de certaines portions d'un ensemble de termes) dans les noms d'ENs soient retrouvées dans les documents **Medline**. Certaines approches [TT04] proposent de générer ces variantes "à la volée", c'est à dire lors de la phase d'extraction des ENs dans les textes, et de mesurer leur similitude orthographique vis-à-vis des noms d'entités références stockés dans les dictionnaires. Néanmoins ce type d'approche présentent deux inconvénients majeurs. D'une part le temps de calcul nécessaire pour traiter chaque instance d'entité découverte dans les textes peut s'avérer très couteux et poser une limite forte à l'exploration de très grands ensembles de documents. D'autre part, ces méthodes ne peuvent utiliser certaines connaissances *a priori* très utiles pour générer des formes variantes spécifiquement retrouvées associées à certaines classes d'objets biologiques. Nous n'utilisons ce type d'information lié à la base de données d'origine, et donc inaccessible lors de l'étape d'extraction des ENs, que dans un cas très précis : la résolution des formes acronymiques.

- Des alias sont générés *de novo* au sein de dictionnaires à partir des combinaisons des différentes formes acronymiques et des noms complets d'une même EN. Par exemple, "Chemokin like receptor 1", "CMKLR1", "CMKL receptor 1", "CMK light R 1", "Chemokin L receptor 1", "CMK light receptor 1", "Chemokin like R 1" et "Chemokine LR 1". En général, certaines de ces formes mixtes sont très couramment retrouvées dans les articles scientifiques alors que d'autres sont relativement

rares ou exotiques, néanmoins les bases de données d'origine ne les proposent pas toutes et ces nouveaux alias sont alors ajoutés dans les dictionnaires. L'ensemble des alias nouvellement générés a une réalité biologique et est terminologiquement valide bien que certains de ces alias puissent ne jamais être vus dans les textes. Selon la banque de données terminologique d'origine et sa nomenclature, nous pouvons déterminer l'ensemble des couples symbole/forme développée admis ou incompatibles pour chaque classe d'ENs. Un symbole est ici défini comme un caractère unique ou une suite de caractères alphabétiques et représentant un terme ou un ensemble de mots (le symbole est souvent une abréviation ou un acronyme à proprement parlé). Un lexique de plus de 40 symboles avec leurs définitions correspondantes, toutes classes d'ENs confondues, est utilisé. Voir le tableau 3.2 pour illustration.
Lorsque nous disposons à la fois d'une forme entièrement ou partiellement acronymique ou abréviée et d'une forme complète du nom d'une même entité à partir d'une banque de données source, nous essayons de développer les lettres et combinaisons de lettres de la forme abréviée grâce à notre lexique si et seulement si le terme développé est présent dans le nom complet et à une position compatible à la fois dans l'acronyme et dans le nom complet. En cherchant à ne mettre uniquement en correspondance des acronymes et des formes développées au sein des ressources terminologiques, nous évitons les ambigüités liées aux définitions multiples de certains acronymes.
Nous parcourons avec une fenêtre glissante de taille de caractères variable l'acronyme de la droite vers la gauche. La définition correspondante à la forme abréviée définie dans notre lexique et spécifiée dans le cadre de cette fenêtre est alors recherchée dans le nom complet. La fenêtre de recherche est initialisée à la taille maximale des symboles présents dans le lexique puis le caractère le plus à gauche est retiré de la fenêtre de recherche si la mise en correspondance est infructueuse et ce jusqu'à ce qu'une définition d'un symbole soit repérée dans le nom complet ou que la fenêtre atteigne une taille nulle. Dans le cas où le symbole n'a pas été résolu, la fenêtre de recherche est décalée vers la gauche de un caractère. Dans le cas contraire, celle ci est décalée de la taille du symbole résolu et ceci afin d'éviter les chevauchements de définitions. L'ensemble des combinaisons symboles/formes complètes est ensuite généré et ajouté dans les dictionnaires.
La figure 3.1 présente l'algorithme correspondant. Les variables et les fonctions utilisées sont définies dans l'encart 3.2.

Par exemple, la séquence de caractères provenant de notre lexique (ici un seul ca-

Symbole	Définition	Classe d'EN correspondante
A	alpha	Gène ou protéine
A	alpha	Site de liaison aux facteurs de transcription
AE	activating element	Site de liaison aux facteurs de transcription
AP	accessory protein	Gène ou protéine
AP	accessory protein	Site de liaison aux facteurs de transcription
AP	associated protein	Gène ou protéine
AP	associated protein	Site de liaison aux facteurs de transcription
AS	activating sequence	Site de liaison aux facteurs de transcription
AS	antisense	Gène ou protéine
B	beta	Gène ou protéine
B	beta	Site de liaison aux facteurs de transcription
B	box	Site de liaison aux facteurs de transcription
BE	binding element	Site de liaison aux facteurs de transcription
BP	binding protein	Gène ou protéine
BP	binding protein	Site de liaison aux facteurs de transcription
BS	binding site	Site de liaison aux facteurs de transcription
C	catalytic	Gène ou protéine
C	catalytic	Site de liaison aux facteurs de transcription
CE	conserved element	Site de liaison aux facteurs de transcription
CL	c terminal like	Gène ou protéine
CR	chromosome region	Gène ou protéine
CS	conserved site	Site de liaison aux facteurs de transcription
D	delta	Gène ou protéine
D	delta	Site de liaison aux facteurs de transcription
D	domain	Gène ou protéine
D	domain	Site de liaison aux facteurs de transcription
DC	domain containing	Gène ou protéine
E	element	Site de liaison aux facteurs de transcription
E	epsilon	Gène ou protéine
E	epsilon	Site de liaison aux facteurs de transcription
FAM	family	Gène ou protéine
G	gamma	Gène ou protéine
G	gamma	Site de liaison aux facteurs de transcription
H	hormone	Gène ou protéine
H	hormone	Site de liaison aux facteurs de transcription
IE	inducible element	Site de liaison aux facteurs de transcription
IN	inhibitor	Gène ou protéine
IP	interacting protein	Gène ou protéine
K	kappa	Gène ou protéine
K	kappa	Site de liaison aux facteurs de transcription
K	kinase	Gène ou protéine
K	kinase	Site de liaison aux facteurs de transcription

Fig. 3.2 – Exemples d'entrées du lexique de mise en correspondance des symboles et de leurs définitions dans le contexte des gènes, protéines ou sites de liaison aux facteurs de transcription

TAB. 3.1 – Génération des formes mixtes symbole/nom complet

```
Pour posAcro de tailleAcro à 1 faire
Début
   Pour tailleFenetre de dico[maxTailleAcro] à dico[minTailleAcro] faire
   Début
      Si posAcro<>tailleFenetre alors Sortir de la boucle

      resolutionOK ← faux

      Pour indiceAcro de dico[tailleFenetre, minIndiceAcro] à dico[tailleDeFenetre, maxIndiceAcro] faire
      Début
         Si Portion(Acronyme, posAcro, tailleFenetre) = dico[tailleFenetre, indiceAcro] alors
         Début
            Pour indiceNomComplet de dico[tailleFenetre, indiceAcro, minIndiceNomComplet] à dico[tailleFenetre, indiceAcro, maxIndiceNomComplet] faire
            Début

               Pour posNomComplet de tailleNomComplet à 1 faire
               Début
                  Pour tailleFenetreNomComplet de (tailleNomComplet - posNomComplet) à 1 faire
                  Début

                     Si Portion(NomComplet, posNomComplet, tailleFenetreNomComplet) = dico[tailleFenetre, indiceAcro, indiceNomComplet] alors
                     Début

                      symbolesResolus[maxIndiceSymboleResolu + 1] ← (dico[tailleFenetre, indiceAcro, indiceNomComplet], posNomComplet)

                      posAcro ← posAcro - tailleFenetre + 1 ;
                      resolutionOK ← vrai
                      Sortir de la boucle

                     Fin Si

                  Fin Pour
                  Si resolutionOK = vrai alors Sortir de la boucle
               Fin Pour
               Si resolutionOK = vrai alors Sortir de la boucle

            Fin Pour
            Si resolutionOK = vrai alors Sortir de la boucle
         Fin Si
         Sirésolutionok = vrai alors Sortir de la boucle
      Fin Pour

   Fin Pour
Fin Pour

listeFormesMixtes ← Combinaisons(nomComplet, symbolesResolus)
```

TAB. 3.2 – Définitions des variables et des fonctions de l'algorithme

`posAcro` et `posNomComplet` sont les positions en cours sur l'acronyme et sur le nom complet respectivement.

`tailleAcro` et `tailleNomComplet` sont les tailles (en nombre de caractères) de l'acronyme et du nom complet respectivement.

`tailleFenetreAcro` et `tailleFenetreNomComplet` sont les tailles (en nombre de caractères) des fenêtres glissantes sur l'acronyme et sur le nom complet respectivement.

`dico` est un tableau multidimensionnel dont la première dimension représente la taille des acronymes inscrits `[maxTailleAcro...minTailleAcro]`, la deuxième dimension les acronymes de la taille mentionnée `[maxIndiceAcro...minIndiceAcro]` et la troisième dimension les définitions (texte complet) correspondantes à l'acronyme `[maxIndiceNomComplet...minIndiceNomComplet]`.

`indiceAcro` et `indiceNomComplet` sont les indices respectifs d'un acronyme et d'un nom complet au sein de `dico`.

`Portion` est un fonction qui extrait une portion du texte. Ses arguments sont la chaîne de caractères qui contient la zone à extraire, l'indice du premier caractère de cette portion et le nombre de caractères à extraire à partir de l'indice spécifié.

`symbolesResolus` est un tableau `[maxIndiceSymboleResolu]` de structure {un acronyme détecté dans un nom complet, la position de cet acronyme sur le nom complet}.

`Combinaisons` est une fonction qui génère l'ensemble des combinaisons des formes hybrides symbole/nom complet. Ses arguments sont la chaîne de caractères du nom complet et le tableau `symbolesResolus` correspondant.

`resolutionOK` est un drapeau spécifiant si un symbole donné a été détecté dans le nom complet.

ractère) "R" et sa définition "receptor" sont retrouvés dans le symbole "CMKLR1" et dans le nom complet associé "Chemokin like receptor 1", respectivement. Deux alias alternatifs sont alors générés, d'une part "CMKL receptor 1" et d'autre part "Chemokin like R 1". De la même manière, le caractère "L" et sa définition "like" serviront à définir deux nouveaux alias à partir des deux alias déjà retrouvés et du symbole et de la forme complète originaux. Il est à noter que si le mot "like" avait été retrouvé à droite du mot "receptor" dans le nom complet ou si le caractère "L" avait été retrouvé à droite du caractère "R" dans le symbole, les nouveaux alias correspondants n'auraient pas été générés. Le développement des formes symboliques et la réduction des noms complets suivent le sens de lecture standardisé des noms d'entités, du plus précis au plus général et de la droite vers la gauche. La figure 3.3 schématise cet exemple particulier.

– Nous créons aussi de nouveaux alias à partir des combinaisons issues des arrangements des termes dits descriptifs lorsqu'ils sont annotés en tant que tel dans les

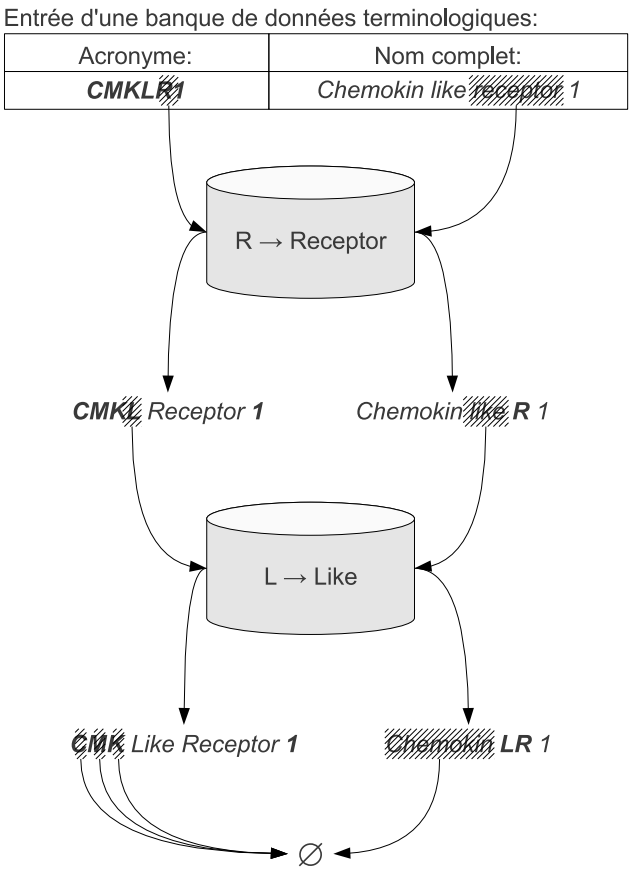


FIG. 3.3 – Génération de formes mixtes acronyme/nom complet à partir du couple "CMKLR1"/"Chemokin like receptor 1"

entrées des ressources terminologiques. Un terme descriptif n'est pas essentiel à la définition de l'EN associée et sa position n'est pas statique dans la chaîne de caractères. Un ensemble de règles établies à partir de la nomenclature propre à chaque source terminologique nous permet de définir les positions autorisées et interdites au sein de l'EN. Par exemple, l'entrée **LocusLink** "Aconitase 1, soluble" donne lieu à la génération des alias "Aconitase 1 soluble", "soluble Aconitase 1", et "Aconitase 1". Le terme "soluble" peut ainsi être placé devant ou derrière le nom de l'EN, ou être omis. Dans le cas de l'entrée "Aconitase (soluble) 1", et selon les recommandations de nomenclature édictées par **LocusLink**, "soluble" est un terme facultatif mais dont la position dans le nom est fixe. Les alias ainsi générés sont au nombre de deux, d'une part "Aconitase 1" et d'autre part "Aconitase soluble 1".

3.1.2.3 Normalisation des noms

A l'étape précédente nous avons répertorié dans nos dictionnaires l'ensemble des variantes syntaxiques complexes et des variantes lexico sémantiques des noms des objets biologiques manipulés. Ces types de variantes ne peuvent être aisément retrouvés *à la volée* à l'étape d'EEN car leur formulation nécessitent certaines formes de connaissances *a priori* uniquement disponibles à partir des bases de données terminologiques. En revanche, d'autres types de variantes sont uniquement basées sur des différences d'écriture liées aux règles orthographiques de l'anglais standard ou de l'anglais scientifiques et ne sont pas spécifiques aux nomenclatures des bases de données terminologiques. A la différences des variantes présentées dans la section précédente, les variantes de type morphologique ou orthographique peuvent être découvertes lors du processus d'EEN sans être préalablement consignées dans nos dictionnaires. Ceci représente un avantage conséquent en terme de simplification des requêtes sur les dictionnaires et de gain d'unité de stockage. En effet, le nombre des variations basées sur des différences de ponctuation, casse de caractères et certaines formes de convention d'écriture par exemple peut être très important. La combinaison de ces différentes formes de variations rend leur enregistrement dans nos dictionnaires problématique.

Nous avons opté pour un approche hybride de gestion des variantes d'ENs. Celle-ci combine l'utilisation des formes variantes hétérogènes présentes dans nos dictionnaires et la normalisation *à la volée* des variations dérivées de la langue anglaise standard ou des conventions d'écriture communes au domaine de la biologie moléculaire, lors de l'étape d'EEN. Ainsi, tous les problèmes de variation des noms ne nécessitant pas, afin de les résoudre, de connaissances exclusives aux ressources terminologiques utilisées sont repoussés à l'étape d'EEN. Afin de limiter le nombre de requêtes à effectuer sur les dictionnaires, nous ne générons pas l'ensemble des variantes orthographiques d'une EN extraite des textes afin d'en retrouver une occurrence au sein des dictionnaires. En revanche, nous transformons chaque entrée des dictionnaires dans un format particulier capable de représenter et d'absorber chaque variation orthographique observée dans les textes. Si les ENs extraites des textes sont elles aussi formalisées grâce à ce mode de représentation, il est possible de limiter le nombre d'opérations nécessaires à identifier l'EN et de n'effectuer qu'une seule requête sur les dictionnaires à la recherche d'une instance de cette forme. L'étape qui consiste en la transcription d'une variante particulière en une représentation à la fois générique et discriminante d'une EN est ici appelée normalisation. La normalisation doit répondre à deux critères impératifs, d'une part elle doit générer une représentation

de l'EN assez générique pour prendre en compte toutes les variations orthographiques ou morphologiques acceptables et d'autre part ce formalisme ne doit pas créer d'ambigüités pouvant conduire à la confusion de deux entités distinctes au sein de nos dictionnaires.

L'étape de normalisation implique l'application séquentielle de différentes règles de transformation :

- Tout d'abord, les ENs sont uniquement conservées sous la forme de noms composés (*compound nouns*) au sein des dictionnaires et transformées au besoin. Par exemple, des trois expressions symbolisant la même entité : "Linker for activation of T cells", "Activation of T cells linker" et "T cells activation linker", seule la dernière est représentée dans nos dictionnaires. L'algorithme prend en charge la transformation des entités à la détection des prépositions 'of', 'in', 'at', 'on', 'by', 'for', 'from', 'to' et 'with'. Ces prépositions forment des points de rupture de la séquence des termes de l'entité et les blocs de termes ainsi délimités sont ré-ordonnés dans le sens inverse de lecture. Lorsqu'un bloc a été déplacé, les articles qui démarre ce bloc doivent être supprimés et les verbes au participe passé qui le terminent sont repositionnés en début de bloc. Par exemple, l'entité "Nuclear factor *activated* **by** T Cell' est enregistrée dans les dictionnaires sous la forme "T Cell *activated* nuclear factor".
- Le problème des majuscules et des minuscules en biologie est très important surtout pour le nom de molécules et des cellules (par exemple, "cAMP" ne correspond pas à la même entité que "CAMP" alors que "CAMP" et "Camp" sont considérés comme étant synonymes). La présence de caractères spéciaux et d'espaces est souvent aussi source d'ambigüité (par exemple, "IL2R", "IL2 R" et "IL2-R" représentent la même entité). Nous utilisons un algorithme qui nous permet à la fois d'absorber les différences de ponctuation ainsi que de différencier les rôles joués par la casse de caractères lorsqu'ils sont significatifs.

  La chaîne de caractères est décomposée en blocs fonctionnels séparés par des espaces au besoin. La délimitation s'effectue à chaque changement de la casse de caractères ou entre un caractère alphabétique et un numéro ou un caractère spécial. Ainsi, chaque bloc ne doit au final contenir que des majuscules, minuscules ou des numériques et caractères spéciaux. Finalement, l'ensemble de l'expression est converti en minuscule et les caractères spéciaux (non alpha-numériques) sont supprimés.

  Une attention particulière est portée à la présence d'une seule majuscule en début de mot. Dans ce cas, la majuscule n'est pas séparée du bloc suivant. En effet, et ici la langue anglaise standard rejoint les règles d'écriture des domaines techniques de manière générale, ceci est considéré comme une information typographique par-

ticulière et non en tant qu'indication pour différencier deux entités distinctes. La conservation de ce type d'indices typographiques (par exemple, l'utilisation de la majuscule afin de différencier la forme gène/protéine d'une même entité) n'a pas lieu d'être à l'étape de stockage des entités dans les dictionnaires.

Par exemple, "cAMP", "c-Amp" et "c Amp" qui symbolisent la même entité sont transformés sous la forme "c amp". "CAMP", une entité différente, est à son tour transformée sous la forme distincte "camp".

Quelques exceptions doivent être prises en compte à cette étape et intégrées en tant que règles particulières au sein du processus de normalisation :

- Certaines formes de capitalisation doivent être traitées à part. C'est le cas notamment des abréviations communes de l'unité de mesure Dalton et ses multiples. Par exemple, "kD", "kDa" et "kd" sont communément retrouvés dans les textes et sont équivalents. D'autre part, les marques du pluriel associées à un acronyme (par exemple, "IL2Rs") ou le cas possessif (par exemple, "Crohn's Disease") sont supprimés du nom de l'entité. Après normalisation, un bloc intermédiaire contenant le caractère unique 's' est caractéristique de ce type d'information purement grammatical.
- Certains caractères spéciaux font partie intégrante du nom de l'entité nommée et ne doivent pas être supprimés. C'est le cas des caractères '-' et '.' lorsqu'ils sont associés à un bloc contenant uniquement des numéros, ici indicateurs d'un chiffre négatif ou d'une décimale. D'autre part, un ensemble de règles spécifiques a été développé afin de rendre compte des particularités des symboles utilisés dans les domaines de la génétique (par exemple, "-/-", symbole pour double mutant négatif) et de la chimie (par exemple, "H+", symbole du proton et "3"' désignant un numéro de carbone sur un cycle).

- Dans le domaine générique de la biologie, certains symboles sont communément acceptés et reconnus, indépendamment de la spécialité technique concernée. C'est le cas des chiffres romains et des symboles grecs qui peuvent être employés, selon les auteurs, tels quels ou encore écrits sous la forme de numériques ou en toutes lettres, respectivement. Nous transformons les chiffres romains isolés dans un bloc en caractères numériques. De façon similaire, nous remplaçons les symboles grecs par leur définition. Encore une fois, quelques exceptions demeurent et doivent être prises en charge. Par exemple, le caractère 'X', précédé ou suivi d'une instance du mot "chromosom" ou du mot "ray", ne désigne pas dans ce cas de figure particulier le numéro 10 et ne doit pas être transformé.
- Finalement, chaque mot appartenant à une variante est racinisée en utilisant l'algo-

rithme de Porter [Por80] à la condition qu'il ne corresponde pas à un acronyme. La racinisation a pour but de retrouver les racines grammaticales des mots et permet de s'affranchir, en particulier, des formes plurielles et de la différence entre les suffixes d'origine américaine ou britannique. La racine ainsi produite doit être identique pour tous les mots ayant un sens commun. Une racinisation forte à comme principal désavantage d'occulter les formes actives et passives des verbes (par exemple, "neutrophil activated peptide" et "neutrophil activating peptide" ne peuvent représenter la même entité biologique). Nous utilisons une version légèrement modifiée de l'algorithme de Porter afin de conserver les différences entre les participes passés et les participes présents ou les gérondifs au sein de nos dictionnaires. Nous avons aussi intégré les règles complémentaires de racinisation proposées par Yamout et al. [YDHS04] au sein de l'algorithme original. Ces modifications permettent de limiter les erreurs issues de la "sur-racinisation", c'est à dire lorsque deux mots distincts sont réduits sous la même racine par erreur, sur certains termes contre-exemples qui ne suivent pas les règles de racinisation classiques.

En étudiant la structure des ENs présentes dans les publications scientifiques, nous avons remarqué que les noms des entités répertoriées dans les bases de données d'origine sont très souvent formels et très descriptifs en comparaison. On a rarement dans les textes la totalité des termes définis dans les bases de données. En conséquence, nous appliquons une dernière règle de normalisation afin de supprimer les termes jugés non informatifs avant le stockage de la variante dans les dictionnaires. Ces règles utilisent une liste prédéfinie de mots clefs supprimables et retrouvés uniquement en rapport avec la précision de l'appartenance à un groupe ou à une famille biologique (par exemple, "precursor", "type" ou member").

3.1.3 Erreurs d'annotation

Le contenu des bases de données terminologiques sélectionnées est soumis à validation et est vérifié par l'organisme responsable de la base de données. Néanmoins, il n'est pas rare que les données enregistrées ne respectent pas entièrement les formalismes d'écriture décrits par les règles de nomenclature de la base de données. Le mode de représentation des noms d'entités et notamment la façon de signaler les termes descriptifs et l'appartenance spécifique à une espèce animale peut ne pas être homogène au sein d'une même base de données. L'information terminologique peut être ainsi mélangée avec des indications fonctionnelles ou des précisions purement illustratives. L'absence de règles de nomencla-

ture strictes ou le non respect de ces règles lors de la soumission des données aux bases de données entrainent le bruitage d'une partie des informations terminologiques proposées. Afin de conserver une fiabilité maximale des dictionnaires créés, seule les données terminologiques doivent être conservées et un travail de nettoyage *a posteriori* des entrées de bases de données est nécessaire. Nous avons répertorié et isolé deux classes d'erreurs couramment retrouvées dans l'annotation des entités nommées des bases de données :

- Tout d'abord, des précisions appartenant à des concepts de plus haut niveau peuvent être incluses. Ces informations sont d'ordre fonctionnel, structural ou peuvent spécifier la super classe biologique dont l'entité fait partie. Par exemple, la mention "transmembrane transporter" dans l'entrée **HUGO** "Dolichol kinase (transmembrane transporter)" n'est pas une information terminologique propre à l'entité biologique "Dolichol kinase" mais une précision sur la structure de la protéine correspondante. Le rôle des parenthèses dans la base de données **HUGO** étant réservé à préciser un alias ou le nom de l'espèce animale où l'entité à été découverte, l'information "transmembrane protein" est incorrecte dans ce contexte et est donc source d'ambigüité.
- Ensuite, des termes sans intérêt (par exemple, "uncharacterized protein") ou parasites (par exemple, "contains only BH3 domain", "antigen identified by monoclonal antibody" ou "entry checked") peuvent être présents. Ces termes ne présentent aucune information utilisable pour une application de REN et sont ajoutés majoritairement afin de spécifier les conditions expérimentales dans lesquelles l'entité a été identifiée pour la première fois.

En complément d'une validation manuelle des données contenues dans les dictionnaires, un ensemble de règles de nettoyage automatique a été développé afin de limiter la tâche de l'expertise humaine. Ces règles, basées sur l'utilisation d'expressions régulières, permettent de supprimer directement des entrées des bases de données les informations non pertinentes à partir des exemples les plus fréquemment retrouvées lors d'une première analyse manuelle des données. Nous avons tiré aléatoirement 1000 entrées des dictionnaires, soit autant de variantes, ce qui correspond approximativement à 0,4% du contenu total, et ce afin d'estimer la qualité d'annotation des dictionnaires. Après analyse manuelle de ces entrées, nous avons comptabilisé environ 4% d'entrée erronnées ou nécéssitant au moins une correction. Ces corrections ont été utilisées en tant que support pour développer les règles de nettoyage automatisé.

3.1.4 Evaluation de la couverture des dictionnaires

Nous avons dans un premier temps évalué la couverture des dictionnaires crées sur un corpus de test basé sur **GENIA**. Nous cherchons à déterminer la proportion d'ENs à la fois présentes dans nos dictionnaires et dans le corpus (vrais positifs et faux négatifs). Le nombre de variantes de nos dictionnaires qui ne correspondent à aucune entité biologique réelle (faux positifs) est analysé dans la section suivante. Les faux positifs correspondent à des entrées erronées de nos dictionnaires et qui n'ont ainsi pu être prises en charge lors de la phase de nettoyage. Leur découverte dans les textes nécessite la mise en œuvre de techniques d'EEN.

Nous avons sélectionné aléatoirement des phrases issues du corpus **GENIA 3.0** que nous avons ré-annotées pour les besoins spécifiques de cette évaluation. **GENIA** est le corpus le plus adapté pour les études d'IEN dans le domaine de la biologie moléculaire et le plus complet. Néanmoins quelques modifications de son contenu doivent être apportées afin de rendre compte des spécificités de notre approche :

- D'une part nous tenons à nous assurer de la qualité des appariements et du recoupement de nos propres classes d'objets biologiques avec celles présentes dans **GENIA**. En effet, la classe **GENIA** générique 'protein' correspond grossièrement à la fusion des nos classes 'protéines' et 'facteurs de transcription'. Les différentes instances des classes 'protein' dans le sous ensemble du corpus **GENIA** ont donc été modifiées en correspondance. De la même manière, la classe **GENIA** 'DNA' peut représenter un 'gène' ou un 'site de liaison aux facteurs de transcription' selon notre propre classification et doit être ré-annotée en conséquence. Il est à noter que **GENIA** distingue les lignées cellulaires ('cell line') des types cellulaires ('cell type'), ce que nous ne faisons pas. Les deux classes sont donc fusionnées.
- D'autre part, nous avons aussi ajouté notre propre annotation lorsque le concept n'était pas préalablement référencé dans **GENIA**. C'est notamment le cas des 'protocoles et appareillages expérimentaux' et des 'tissus ou organes'.

La séquence de mots correspondant à une EN d'une des classes d'intérêt est isolée du reste du document et associée à sa classe d'appartenance. Le corpus de test utilisé afin de quantifier la couverture de nos dictionnaires contient 200 instances d'EN annotées.

Chaque EN est alors normalisée et son occurrence est recherchée au sein de nos dictionnaires. Dans nos résultats nous comptabilisons en tant que vrais positifs les ENs à la fois présentes dans nos dictionnaires et dans les textes sous certaines conditions. La

variante trouvée dans les textes doit être strictement identique à celle présente dans les dictionnaires. D'autre part, une des classes enregistrées dans les dictionnaires pour l'entité détectée doit être équivalente à celle de l'instance dans le sous corpus **GENIA**. Dans le cas où la variante détectée dans les textes peut correspondre à plusieurs entités de classes différentes dans les dictionnaires, nous ne cherchons pas à identifier la classe effective parmi les choix correspondants. Seule la présence où l'absence d'une entité compatible est ici mesurée. La mise en œuvre et l'évaluation des techniques de désambiguïsation nécessaires pour répondre à cette difficulté sont présentés dans le chapitre suivant. De la même manière, nous ne cherchons pas à distinguer les entités appartenant à l'humain des entités originaires de mammifères. Toute variante présente dans les textes mais absent des dictionnaires est considérée en tant que faux positif. Si la variante est bien détectée au sein de nos dictionnaires mais qu'aucune entité associée n'est présente sous la classe définie par l'instance dans le sous corpus **GENIA**, nous le considérons aussi en tant que faux négatif.

La composition des dictionnaires et les résultats de l'évaluation de leur couverture sur le corpus de test sont présentés dans le tableau 3.3.

TAB. 3.3 – Composition des dictionnaires et couverture sur le corpus de test

	G	S	F	C	O	P	TOTAL
Composition							
Nombre de variantes	183196	6524	11379	508	1768	1284	204659
Nombre d'entités	43259	2268	1773	312	1202	769	49583
Couverture							
Nombre d'entités annotées	50	9	40	40	40	20	200
Nombre de vrais positifs	44	7	34	37	32	18	167
Nombre de faux négatifs	6	2	6	3	8	2	33
Rappel	0,88	0,78	0,85	0,93	0,8	0,9	0,84

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

Les erreurs sont soit dues à l'absence de la forme variante correspondante dans nos dictionnaires soit parce que l'entité n'a pas du tout été référencée à l'origine dans les bases de données.

3.1.5 Résumé

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

scientifiques du domaine de la biologie moléculaire. Afin de simplifier la tâche de REN, nous déléguons la gestion des problèmes liés à la variété des formes des ENs à cette étape et non lors de l'EEN. Le principal facteur influant sur la qualité de tels dictionnaires est le nettoyage systématique et expertisé des données redondantes, inappropriées ou erronées. Un soin particulier a été apporté à l'automatisation de la tâche de nettoyage, néanmoins elle ne peut se substituer complètement à l'expertise manuelle. En conséquence, la mise à jour automatique de dictionnaires prêts à l'emploi n'est pas encore réalisable entièrement, même si les procédures mises en place constituent une aide précieuse.

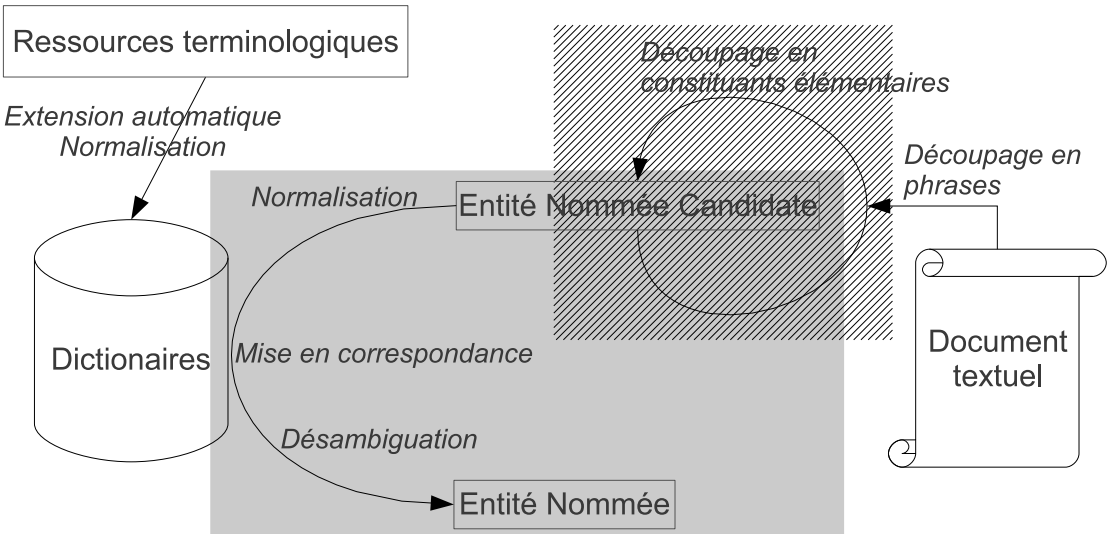
3.2 EEN et IEN

Cette section présente une méthode d'EEN et d'IEN en biologie moléculaire. Les ENs peuvent être de différentes classes et nous ne nous limitons pas à la reconnaissance de gènes et de protéines uniquement. Cette méthode repose sur l'extraction de syntagmes nominaux en tant qu'ENs candidates à partir de règles syntaxiques. Ces ENs candidates sont ensuite identifiées ou rejetées grâce à l'utilisation de dictionnaires développés dans la section précédente. Le couplage des méthodes d'EEN et de l'utilisation de dictionnaires a été montré comme étant une solution efficace à la question de l'IEN en biologie moléculaire [KKT03]. Nous discutons dans cette section des techniques mises en œuvre afin de répondre aux problèmes de détection des bornes d'une EN dans les textes et des solutions de désambiguïsation de l'identité des ENs extraites. Les processus d'EEN et d'IEN sont schématiquement proposés dans la figure 3.4. Les performances de notre système d'EEN sont mesurées grâce au corpus de test utilisé lors de la compétition **BioNLP/NLPBA 2004** et sont comparées aux résultats obtenus par les participants au concours. La tâche d'IEN est elle évaluée sur un sous-ensemble de ce corpus de test que nous avons enrichi.

3.2.1 EEN

Nous utilisons des règles grammaticales simples afin de ne considérer uniquement que les groupes de mots pouvant potentiellement représenter des ENs du domaine de la biologie moléculaire au sein des textes des publications à analyser. Ces mots ou groupes de mots sont ensuite confrontés avec les entrées de nos dictionnaires à la recherche de correspondances. Cette stratégie nous permet d'éviter d'explorer chaque portion de texte à la recherche d'ENs, et donc de restreindre les temps de calcul nécessaires, tout en ayant la

Fig. 3.4 – Processus d'EEN (zone hachurée) et d'IEN (zone grisée)



Les deux processus sont étroitement inter-connectés et partagent des méthodes communes.

possibilité de découvrir les ENs aux formes les plus élaborées et les plus complexes (voir la section 2.4.2.3).

3.2.1.1 Pré-traitements

Lors de cette étape d'EEN, nous nous positionnons à l'échelle de la phrase. La phrase possède une unité sémantique (ou unité de communication), c'est-à-dire, un contenu transmis par le message (sens, signification...). Ce contenu se dégage du rapport établi entre les signes de la phrase, et dépend du contexte et de la situation du discours : chaque phrase a sa référence. Cette référence résulte de la mise en rapport avec une situation, même imaginaire, de discours. Le mot seul n'est rien. Il ne se définit que par rapport aux autres éléments de la phrase. Il faut remarquer que le sens ne dépend pas seulement des mots (aspect lexical). L'organisation grammaticale y est aussi très importante : c'est l'aspect syntaxique. Normalement, la syntaxe ne dépasse jamais les limites de la phrase.

Nous extrayons et isolons donc dans un premier temps les phrases des textes scientifiques d'intérêt. Nous utilisons les règles proposées par Mikheev [Mik99] que nous avons adaptées au domaine biomédical. Nous recherchons séquentiellement dans les textes les marqueurs potentiels des limites des phrases que sont '.', '?', '!', '...'. Ces marqueurs sont

automatiquement isolés des termes lorsqu'ils les terminent et sont donc désormais considérés comme étant des mots à part entière. Par exemple, dans la portion de texte "(...)but activates IL2. However it has not(...)" le point est séparé du terme IL2 et l'expression est transformée en "(...)but activates IL2 . However it has not(...)". Lorsqu'un tel signe de ponctuation est détecté nous isolons en même temps le contexte local de son apparition, c'est à dire les mots directement adjacents au marqueur et ce afin de déterminer si sa présence indique la fin d'une phrase ou non. Le contexte local est défini de manière très restreinte par les deux mots qui précèdent la ponctuation ainsi que les deux mots qui le suivent. Un *POS* est alors attaché à chacun de ces mots, selon la méthode décrite dans le paragraphe ci-dessous, ainsi qu'une catégorie morpho-grammaticale particulière (voir le tableau 3.4).

TAB. 3.4 – Catégories morpho-grammaticales utilisées pour la détection des fins de phrase

Catégorie	Description
NONE	joker utilisé pour signaliser l'absence de mot à la position spécifiée
ABBR	le mot appartient à une liste contrôlée d'abréviations ou de mots réservés se terminant par un signe de ponctuation (par exemple, "Approx", "Ssp", "U.V" et "i.e")
MORPHOSPEC	nom commun qui n'appartient pas aux catégories LOW_COMMON et CAP_COMMON (par exemple, "cDNA", "1,4-diacetyl-1,4-dihydro-4-phenylpyridine" et "PBS")
CLOSING_PUNCT	]) }
OPENING_PUNCT	[({
DISCOURSE_PUNCT	" '
PUNCT	. . . . ; ! ?
LOW_COMMON	nom commun composé uniquement de caractères alphabétiques et écrit en minuscule
CAP_COMMON	nom commun composé uniquement de caractères alphabétiques et écrit en minuscule à l'exception du premier caractère qui est une majuscule
PROPER_NAME	nom propre
LOW_OTHER	n'importe quel autre mot sans majuscule au premier caractère
CAP_OTHER	n'importe quel autre mot avec majuscule au premier caractère
NUM	numérique

Nous considérons par défaut que la présence d'un marqueur de ponctuation est significatif de la fin d'une phrase sauf si l'une des quatre conditions suivantes est vérifiée (n correspond à la position relative du mot par rapport à *mot[n]* qui est le marqueur potentiel de fin de phrase) :

1. mot[n+1] $\in$ (LOW_OTHER $\vee$ LOW_COMMON)
2. mot[n-1] $\notin$ NONE $\wedge$ mot[n] = '.' $\wedge$ mot[n+1] $\in$ PUNCT
3. mot[n-2] $\in$ MORPHOSEC $\wedge$ mot[n-1] $\in$ NUM $\wedge$ mot[n] = '.' $\wedge$ mot[n+1] $\notin$ MORPHOSEC

4. mot[n-1] $\wedge$ ABBR $\wedge$ mot[n] = '.'

Nous avons sélectionné aléatoirement 100 phrases découpées à l'aide de l'algorithme à partir du corpus développé dans la section 3.1.4. Seules deux phrases se sont révélées incorrectement sectionnées. La première erreur est due à la présence de l'abréviation "tabl." pour tableau. La deuxième est liée à l'inclusion de texte explicatif entre parenthèses. Dans la portion de phrase "(...) during our experiment (over night. The samples are randomly pooled.) As a result (...)" la présence du point après le mot "night" conduit au découpage de la phrase de manière incorrecte. En conséquence, nous avons ajouté une règle particulière afin de ne pas prendre en compte les signes de ponctuation de fin de phrase au sein des textes entre accolades, parenthèses, chevrons, crochets, guillemets, apostrophes ou traits d'union, à l'exclusion de ceux situés en dernière position dans l'expression ainsi délimitée. L'usage préconise en effet de placer le point final d'une phrase devant une parenthèse fermante et non après si la phrase se termine par du texte entre parenthèses.

Chacune des phrases alors extraites subit ensuite une première analyse grammaticale dite de surface. Un *POS* est associé à chaque mot de la phrase selon sa fonction grammaticale. Le processus d'extraction des syntagmes nominaux correspondants aux ENs repose sur l'utilisation des *POS* uniquement. Nous utilisons pour lier un *POS* particulier à un mot l'étiqueteur **Genia Part Of Speech Tagger** [TTJD+05]. Cet étiqueteur grammatical de surface a la spécificité d'être particulièrement adapté à la fois aux textes biomédicaux et aux textes généraux de type journalistique. L'algorithme utilisé par l'outil est basé sur les travaux de "réseaux de dépendance cyclique" (*Cyclic Dependency Network*) de Toutanova *et al.* [TKMS03]. Les travaux de Toutanova rapportent un des étiquetages de *POS* en langue anglaise les plus fiables à ce jour sur les articles du **Wall Street Journal**. L'implémentation de l'algorithme au sein de **Genia Part Of Speech Tagger**, et après entrainement en combinant 90% des données des corpus **Wall Street Journal**, **GENIA 3.02** et **PennBioIE 0.9**[5], obtient une précision globale de 98,35% sur le contenu non utilisé de ces corpus. Sur un corpus de 1835 mots, collecté à partir d'articles biomédicaux récents, **Genia Part Of Speech Tagger** réalise un score de précision de 95,1% ce qui le place parmi les meilleurs étiquetteurs du domaine biomédical.

[5] http ://bioie.ldc.upenn.edu/publications/latest_release/

3.2.1.2 Grammaire d'ENs

Afin de discerner des textes les syntagmes contenant potentiellement les ENs d'intérêt nous ne réalisons pas d'analyse syntaxique dite de profondeur. L'utilisation d'un analyseur syntaxique complet annulerait le gain en temps de calcul apporté par la mise en correspondance de portions sélectives de textes avec le contenu des dictionnaires. A l'opposé, nous nous efforçons de développer une grammaire d'ENs basée uniquement sur les *POS* et l'analyse déterministe et non-récursive de la phrase. Seuls des ilots sélectifs de texte sont traités, la compréhension globale de la syntaxe de la phrase n'est pas envisagée. Le but étant de pouvoir analyser des millions de lignes de textes en un temps raisonnable. Cette grammaire doit néanmoins être robuste et être adaptée aux ENs rencontrées dans les textes de biologie (voir la section 2.4.2.3). Il est à noter que la grammaire d'ENs ainsi générée ne doit pas être vue comme un modèle syntaxique rigoureux mais comme une approche heuristique ciblée. Ainsi, les motifs écrits ne sont pas forcément des syntagmes au sens traditionnel du terme, ils sont élaborés dans le but d'être des indicateurs fiables de la structure syntaxique sous-jacente.

Notre méthode s'inspire des travaux d'Abney [Abn96] sur les "cascades à état fini". Une cascade à état fini consiste en une séquence de niveaux. Les syntagmes d'un niveau particulier sont construits à partir des syntagmes du niveau précédent. Il n'y a pas de récursion : les syntagmes ne contiennent pas d'autres syntagmes de niveaux équivalents ou supérieurs. Le niveau de base est celui qui correspond aux *POS*. Les niveaux supérieurs sont composés des syntagmes 'fondamentaux' que sont `NP`, `VP`, `PP`, `AP` et `AdvP` (la liste des *POS* compatibles **Penn Treebank** est présentée dans l'annexe 6). Dans une cascade à état fini, les éléments d'entrée d'un niveau sont réduits à un élément unique à chaque transition. Une illustration de ce mécanisme est proposée dans la cascade suivante :
$N_0 \rightarrow N_1$: NP $\rightarrow$ D ? N* N
$N_1 \rightarrow N_2$: PP $\rightarrow$ P NP
Où $N_x \rightarrow N_{x+1}$ correspond à la transition d'un niveau x vers le niveau directement supérieur. Chaque transition est définie par un ensemble de règles. Les règles proposées sont déterministes dans le sens où à partir d'un même ensemble d'éléments pour un niveau donné, le contenu du niveau directement supérieur est toujours identique et unique. L'ordre des transitions est déterminé à partir de l'étude des ENs présentes dans les dictionnaires et notamment la construction des noms d'entités hybrides à partir de deux ENs différentes ou plus.

Dans les paragraphes suivants nous présentons les différentes règles adoptées afin de découvrir les différents syntagmes d'ENs à partir des phrases.

Pré-traitement des énumérations simples Avant de chercher les ilots d'ENs au sein des phrases, une première étape consiste à ré-assembler les fragments de textes qui dépendent d'une même conjonction de coordination et ce dans certaines situations limitées et spécifiques. Ce pré-traitement est issu de l'observation de l'utilisation de certaines formes d'énumérations impliquant des ENs et telles que retrouvées dans les textes de biologie moléculaire. Il est classique, lorsqu'un auteur souhaite lister une suite d'objets biologiques appartenant à la même famille, de ne préciser que l'identifiant spécifique à l'EN et non son nom complet dès la deuxième position de l'énumération. Par exemple, dans les énumérations suivantes "IL-2 and -6" et "Interferon α, β, or γ", qui utilisent les conjonctions de coordination 'and', 'or' et les virgules, nous ré-écrivons ces expressions sous la forme "IL-2 and **IL**-6" et "Interferon α, **Interferon** β, or **Interferon** γ" respectivement. Cette ré-écriture permet de ré-associer la *tête* de l'énumération à chaque membre de la liste. Nous ne gérons que les cas d'énumération les plus simples :

- les seuls séparateurs des objets de l'énumération reconnus sont les conjonctions de coordination 'and' et 'or' ainsi que les virgules.
- les seuls objets de l'énumération qui sont ré-associables avec une *tête* sont les chiffres romains ou numériques, les lettres grecques (par exemple, 'β') ou leurs définitions (par exemple, 'alpha' pour 'α') ou les caractères alphabétiques uniques (par exemple, 'A' ou 'B').
- Si un des objets d'une même liste ne répond pas aux critères de type décrits ci-dessus, le traitement de l'énumération en cours est arrêté à cette position. Dans le cas où cet objet particulier correspond à une *tête* suivi d'un identifiant, une nouvelle énumération, indépendante de la première, doit alors être prise en compte à partir de cette position. Par exemple, dans la liste suivante "IL-2, -3, and TNF-α, or Interferon α and β", nous isolons trois segments au sein de l'énumération (soulignés dans l'exemple), deux des trois portions correspondent en fait à des sous-listes autonomes possédant leur propre *tête*. L'énumération est transformée en "IL-2, IL-3 and TNF-α, or Interferon α and Interferon β".

Le rôle des *queues* dans les énumérations est ambigü et dépend du contexte. Par exemple, dans la liste suivante "IL2 and IL3 receptor", où "receptor" est la *queue*, nous ne savons pas *a priori* si l'expression est équivalente à "IL2 receptor and IL3 receptor". Pour cette raison nous ne cherchons pas à expliciter l'assignation des *queues* aux différents objets manipulés dans les énumérations.

Notre grammaire d'ENs est construite à partir de quatre niveaux successifs qui sont présentés dans les parties ci-dessous. Le premier niveau consiste en la détection des groupes nominaux de la phrase, le deuxième niveau est spécifique à la découverte des appositions des groupes nominaux, les niveaux trois et quatre sont utiles à la mise en relation des différents blocs extraits soit par le biais des verbes soit grâce à des prépositions et conjonctions, respectivement. Le tableau 3.5 présente de manière synthétiquement les différents niveaux de notre grammaire d'ENs et résume les différents points évoqués ci-dessous.

Niveau 1 : détection des squelettes des ENs. Les points d'entrées de détection des syntagmes d'ENs au sein des phrases sont les ilots de groupes nominaux. Les ENs d'intérêt sont exclusivement basés sur des groupes nominaux qui correspondent à l'épine dorsale (ou squelette) des nom d'entités et sur lesquels viennent se greffer différents autres groupements. Cette information a été confirmée par l'étude des entrées de nos dictionnaires. Nos dictionnaires ne contiennent, par exemple, aucune entrée centrée sur des verbes.

- Nous isolons des phrases les noms communs et propres, grâce aux *POS*, qui constituent les squelettes des ENs. Par exemple, "β adrenergic receptor type 1 identified by macro array".
- Les caractères numériques et les symboles non alphanumériques sont associés à ces blocs s'ils sont positionnés directement à la suite d'un fragment dit squelette. Par exemple, "β adrenergic receptor type 1 identified by macro array".
- Deux fragments sont ensuite automatiquement fusionnés s'ils sont juxtaposés ou uniquement séparés par des adjectifs. Par exemple, "β adrenergic receptor type 1 identified by macro array".

Ces squelettes à base de noms correspondent ainsi au cœur de la dénomination de l'entité biologique.

Niveau 2 : détection des préfixes des blocs cœurs. Certains termes d'un même groupe nominal peuvent faire partie intégrante de l'intitulé de l'EN, d'autres en revanche sont soit facultatifs soit sans rapport avec l'EN.

Certains de ces termes satellites sont rattachés au nom de l'entité en tant que préfixes, en apposition. La réunion d'un groupe cœur et d'un préfixe constitue ainsi un bloc de niveau 2. Les séquences ininterrompues d'adjectifs, numériques et symboles non alphanumériques situées devant un bloc cœur sont étiquetées en tant que préfixes. Chacun de ces termes étant considéré comme autant de blocs préfixes autonomes. Par exemple, dans l'expression "modulated by $\frac{activated}{prefixe}$ $\frac{Interleukin\ 2}{coeur}$" "activated" est un préfixe du squelette

"Interleukin 2".

Note sur la difficulté effective de discerner un terme satellite d'un terme cœur. Les préfixes pris en compte sont uniquement détectés grâce à leur *POS*. Ceci constitue le cas de figure d'apposition la plus simple. Néanmoins, de part la longueur conséquente des groupes nominaux dans le domaine de la biologie, en pratique la localisation des termes satellites est beaucoup plus complexe et ambigüe. Certaines formes d'apposition ne peuvent pas être détectées de cette façon. Les blocs cœurs n'étant composés uniquement que de noms et de symboles, les préfixes (qui sont ici équivalents à la fonction de nom adjectival) et les suffixes (qui sont cette fois des appositions en fin de syntagme) ne sont pas détectés en leur sein. Certains noms correspondent à une action dont l'objet est l'EN (par exemple, "assimilation", "transcription", "screening"), d'autres noms permettent encore d'aider à caractériser l'EN (par exemple, "gene", "protein", "experiment"). Pierre Larrivée [Lar04] nomment ces termes satellites substantifs pré- ou postposés et propose de les classer en quatre catégories principales :

- la coordination (par exemple, "the Mad Max couple" où "Mad" et "Max" sont deux objets biologiques aux fonctions complémentaires),
- la qualification (un nom remplace un adjectif qualificatif),
- la complémentation (par exemple, "IL2 screening" ou "IL2 transcription" où la combinaison des deux noms définit une nouvelle EN dont la classe est absente de nos dictionnaires),
- l'identification (par exemple, "IL2 protein" ou "IL2 gene").

En aucun cas les termes utilisés dans ces contextes particuliers ne sont des composants intrinsèques des noms d'entités. Toutefois le recours à un lexique afin de distinguer les termes satellites des termes appartenant des noms des entités se révèle inefficace : selon l'EN un même terme peut indépendamment être un composant fondamental du nom de l'entité ou non. Si l'on s'intéresse aux deux groupes cœurs suivants, "crystal/NN protein/NN elastase/NN" et "IL2/NN receptor/NN activation/NN", nous observons que les noms des entités d'intérêt soulignés dans les exemples peuvent être respectivement rejetés en fin ou en début de syntagme, les autres mots doivent donc considérés comme des préfixes ou des suffixes. L'analyse de ce schéma exemple NN_1 NN_2 NN_3 présuppose un choix entre deux découpages possibles de cette relation dîtes de prémodification : soit NN_1 prémodifie NN_2 NN_3 (par exemple, "placebo control group") soit NN_1 NN_2 prémodifie NN_3 (par exemple, "bone marrow transplant"). Le patron syntaxique ADJ NN_1 NN_2 pose intrinsèquement des problèmes de décodage similaires à ceux des groupes nominaux de schéma NN_1 NN_2 NN_3. La position des termes au sein d'un syntagme constitué du

même *POS* ou non n'est pas un critère de discrimination des termes satellites.

La gestion des préfixes et des suffixes internes aux blocs cœurs est déléguée à l'étape d'identification (voir le paragraphe 3.2.2).

Niveau 3 : détection des verbes structurants. Il est intéressant de noter que certains blocs préfixes peuvent contenir des ENs indépendantes de celles des squelettes. Dans l'exemple suivant "the $\frac{embryonic}{prefixe}$ $\frac{stem}{prefixe}$ $\frac{cell}{prefixe}$ $\frac{expressed}{prefixe}$ $\frac{RAS}{coeur}$ is a protein", l'expression "stem cell" possède la fonction grammaticale d'un complément du nom et est donc à juste titre considéré comme un préfixe de "RAS". Néanmoins, et comme souligné dans la section 2.4.2.3 et dans le paragraphe ci-dessus, ici "stem cell" et "RAS" sont deux entités biologiques distinctes utilisées de paire dans l'expression précédente afin de définir une troisième entité autonome. Après étude du contenu de nos dictionnaires, nous avons observé que certaines ENs hybrides possèdent la caractéristique commune d'être articulées autour d'un participe passé ou présent. Nous décidons de marquer la présence d'ENs au sein d'autres ENs dans les blocs préfixes en considérant la section à gauche des verbes. Le verbe est lui aussi marqué d'une étiquette particulière : verbe de jonction. L'exemple précédent est désormais étiqueté ainsi : "the $\frac{embryonic}{prefixe}$ $\frac{stem\ cell}{coeur}$ $\frac{expressed}{verbe\_de\_jonction}$ $\frac{RAS}{coeur}$ is a protein".

Dans d'autres situations, les participes passés ou présents peuvent être étiquetés en tant que tels et néanmoins faire partie intégrante du nom d'une EN. Les raisons de ce marquage grammatical différentié des verbes peuvent être variées et dépendre notamment du contexte syntaxique, parfois ambigu, de la phrase ou d'erreurs ponctuelles de l'étiqueteur de *POS*. Nous marquons les verbes au participe passé ou participe présent de la même façon que précédemment s'il permettent de faire la jonction entre un bloc cœur et un bloc préfixe ou un autre bloc cœur. Certaines ENs présentent un verbe au participe passé ou présent en tout première position de leur nom (par exemple, dans l'expression "The detected protein is associated angio-migratory cell protein" le nom d'une protéine débute par le verbe "associated"), ces verbes sont étiquetés en tant que préfixes s'ils précédent un bloc cœur ou préfixe, et ce quelque soit leur *POS*, et si le mot précédent n'est pas un modal (par exemple, 'may' ou can'), un adverbe ou un pronom. Les verbes sont aussi marqués s'ils permettent de réunir un bloc cœur avec des chiffres ou des symboles non alphanumériques non préalablement étiquetés ; dans ce cas, ces derniers sont marqués comme étant des préfixes ou des suffixes selon leur position relative au groupe cœur d'origine.

Niveau 4 : détection des prépositions et conjonctions de subordination de structure. De

très nombreuses ENs contiennent des prépositions de conjonctions de subordination en leur sein. Ils sont souvent utilisés, tout comme les verbes au participe passé et présent, en tant que liant entre deux ENs distinctes afin de construire une supra-entité indépendante. Par exemple, "Cell-adhesion-molecule related/downregulated **by** oncogenes" et "Regulator **of** G-protein Signaling 4" sont des noms valides d'entités biologiques. A ce titre, nous marquons toute conjonction de subordination ou préposition 'of', 'in', 'at', 'on', 'by', 'for', 'to', 'with' séparant deux blocs de niveau 3 et inférieur à cette étape.

Niveau 5 : détection des conjonctions de coordination de structure. Certaines ENs telles que, par exemple, la protéine "The activation of signal transducer **and** activator of transcription" peuvent inclure des conjonctions de coordination 'and' et 'or'. De manière identique à l'étape précédente, de telles conjonctions, lorsqu'elles sont positionnées entre deux blocs de niveau 4 ou inférieur, sont marquées et intégrées dans le nom de l'entité.

3.2.2 IEN

A ce stade, chaque phrase est représentée par un ensemble fini de syntagmes d'ENs candidates, basés sur des groupes nominaux étendus et de complexité variable. Nous partons ainsi de l'hypothèse que chacun de ces syntagmes représente zéro ou une EN. Chaque EN candidate est donc *a priori* constituée d'un bloc unique de niveau 1, 2, 3, 4 ou 5 et issu de couches successives de blocs de niveaux identiques.

Deux difficultés doivent être prises en compte lors de l'identification des ENs à partir des syntagmes extraits :

- Il est possible qu'un syntagme contiennent plus d'une EN. Comme exposé dans la section 2.4.2.3 la combinaison de deux ENs ou plus peut servir de base au nom d'une autre EN, distincte. Dans ce cas, nous cherchons à identifier la méta-entité la plus complexe en priorité. D'une part, les blocs indépendants qui constituent le syntagme d'un niveau 4 ou 5 peuvent contenir chacun une entité. Par exemple, le syntagme de niveau 4 "T cell receptor of T cell clone 1C6" est constitué de l'union des deux blocs "T cell receptor" et "T cell clone 1C6" qui représentent chacun une EN différente. D'autre part, plusieurs ENs peuvent coexister en apposition au sein d'un même bloc de niveau 3 et inférieur. Par exemple, un bloc de niveau 1 tel que "Protein Crystal Growth Elastase" contient deux ENs distinctes, la première représente le nom d'une technique expérimentale alors que la deuxième est le nom d'une protéine.

TAB. 3.5 – Les niveaux de la grammaire d'ENs

Transition	Forme succincte	Exemple
$N_0 \rightarrow N_1$	$E_1 \rightarrow$ (NN|NNS|NNP|NNS)* (NN|NNS|NNP|NNS) (SYM|CD)*	Interleukin Receptor α
$N_1 \rightarrow N_2$	$E_2 \rightarrow$ (JJ|CD|SYM)* E_1	Placental lactogen
$N_2 \rightarrow N_3$	$E_3 \rightarrow$ (E_2 ?|(CD|SYM)*) (VBG|VBN)* E_2 (VBG|VBN)* ((CD|SYM)*|E_2 ?)	Liver expressed antimicrobial peptide 2
$N_3 \rightarrow N_4$	$E_4 \rightarrow E_3$ (IN|TO) E_3	Nuclear factor of activated T cell
$N_4 \rightarrow N_5$	$E_5 \rightarrow E_4$ CC E_4	Ankyrin repeat and BTB domain containing 1

E_x représente l'EN candidate, les formes décrites suivent le formalisme des expressions régulières.
La liste des *POS* utilisés est rappelée dans l'annexe 6.

– Nous identifions une EN lorsque le texte qui lui correspond est identique au texte d'une entrée d'un dictionnaire. Ainsi, il est impératif de chercher à retrouver les bornes exactes du nom de l'entité. Comme évoqué dans la section précédente, nous devons déterminer précisément où se termine un préfixe et où débute un suffixe par rapport à l'EN.

3.2.2.1 Le processus d'IEN.

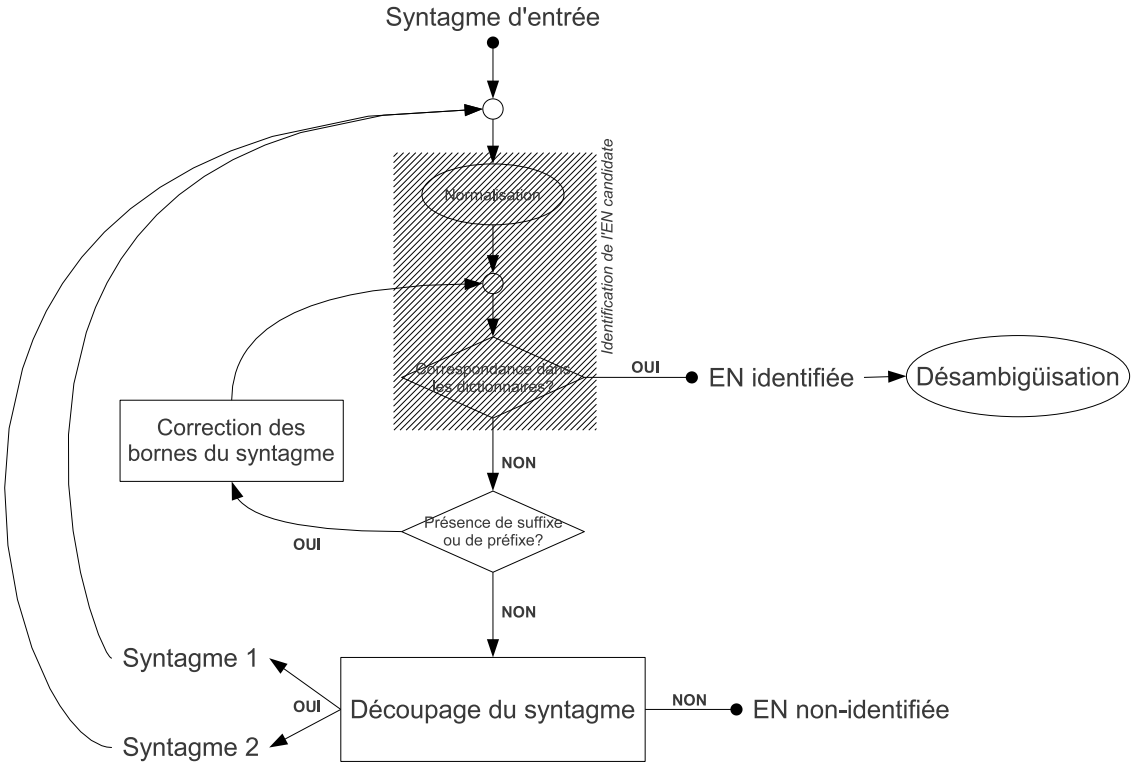
Le processus d'identification, illustré par la figure 3.5, est défini comme suit :

Chaque instance d'EN candidate est cherchée séparément au sein de nos dictionnaires, après normalisation (voir le paragraphe 3.1.2.3). Deux cas de figure sont alors envisageables :

– Si la mise en correspondance du texte avec une entrée des dictionnaires est concluante, une EN a été potentiellement détectée et identifiée. Une phase de désambiguïsation est encore nécessaire afin de s'assurer de la légitimité de l'identité de l'entité extraite (la procédure de désambigüisation sera décrite plus loin).
– Si aucune entrée du dictionnaire ne correspond à l'EN candidate, nous cherchons alors à déterminer les bornes exactes de l'EN candidate au sein du syntagme avant de confronter à nouveau le texte extrait avec le contenu des dictionnaires. En effet, comme évoqué précédemment, l'EN réelle peut être dissimulée à l'intérieur du bloc extrait et doit être isolée précisément afin de ne soumettre que le texte utile à la recherche dans les dictionnaires.
  Deux types de traitements sont alors réalisés sur le syntagme. La deuxième procédure n'est mise en œuvre que lorsque la première ne permet pas l'identification d'une EN. Ces traitements permettent d'exciser au besoin les blocs de différents niveaux afin de mettre à nu une EN selon les règles définies ci-dessous :
  – Dans un premier temps, les bornes du syntagmes correspondant à des blocs satellites sont évaluées :

    1. Si le syntagme débute par un bloc de niveau 2 et contient un préfixe (soit étiqueté en tant que tel ou faisant partie du bloc cœur) ou si le bloc commence par un bloc de niveau 3, cette portion de texte est supprimée et ce afin de générer un nouveau syntagme. Cette opération de suppression du préfixe est réitérée, récursivement, jusqu'à ce qu'une EN soit identifiée ou qu'aucun préfixe ne soit plus disponible. Dans ce dernier cas de figure, les suffixes sont

FIG. 3.5 – Processus d'identification d'une EN candidate



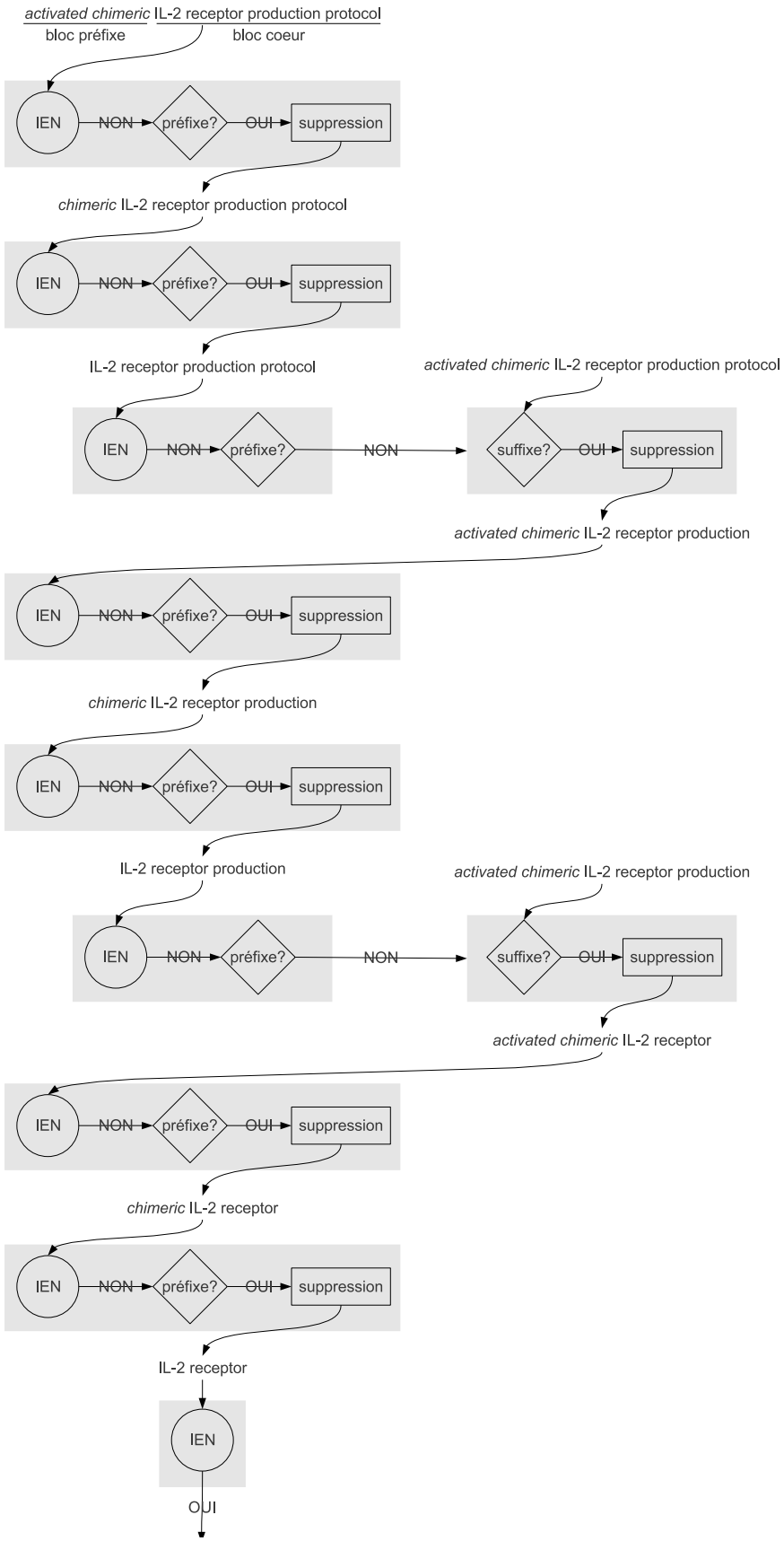
à leur tour testés à partir du syntagme d'origine (le syntagme qui contient tous les préfixes supprimés à cette étape).

2. Si le syntagme se termine par un bloc de niveau 2 et contient un suffixe, le mot et divers symboles associés le plus à droite est éliminé afin de générer un nouveau syntagme. Ce nouveau syntagme est testé contre les dictionnaires et si la mise en correspondance s'avère infructueuse les préfixes de ce syntagme sont alors élagués à leur tour. Lorsqu'aucun suffixe n'est disponible, l'opération de recherche des bornes de l'EN candidate échoue.

La figure 3.6 illustre la méthode de détection des bornes de l'EN candidate à partir du bloc exemple "activated chimeric IL-2 receptor production protocol" où "IL-2 receptor" est une EN.

Nous choisissons de donner volontairement une priorité plus importante aux préfixes et non aux suffixes dans la détermination du nom d'une entité. L'étude des ENs à partir des résultats en section 3.1.4 tend à montrer que les termes satellites sont majoritairement retrouvés sous le forme de suffixes et non de préfixes. Il est à noter que quelques rares cas de construction complexe d'EN tel que "interleukin protein 2", qui correspond dans nos dictionnaires à l'entité "interleukin 2", ne

FIG. 3.6 – Exemple de détection des bornes d'une EN candidate



peuvent pas être traités pour le moment.

La présence de deux ENs en apposition est un problème relativement courant que nous nous devons de gérer. Lorsqu'une EN a été détectée, nous re-soumettons le portion du texte située à sa gauche à la recherche des bornes d'une nouvelle EN indépendante. Dans le syntagme exemple "T-Cell IL-2 Receptor", aucune EN ne correspond à l'intégralité de l'expression. En revanche "IL-2 Receptor" a été correctement identifié comme étant une EN. Si nous devions stopper l'analyse du syntagme à ce ce stade, l'entité "T-Cell", indépendante de "Il-2 Receptor", ne pourrait être découverte. Cette opération n'est pas présentée dans la figure 3.5 dans un souci de concision.

- Si à ce stade aucune EN n'a pu être détectée à partir des dictionnaires, nous considérons désormais que le syntagme d'origine contient plus d'une EN (non disposées en apposition). Chaque bloc constituant le syntagme peut représenter une EN autonome.

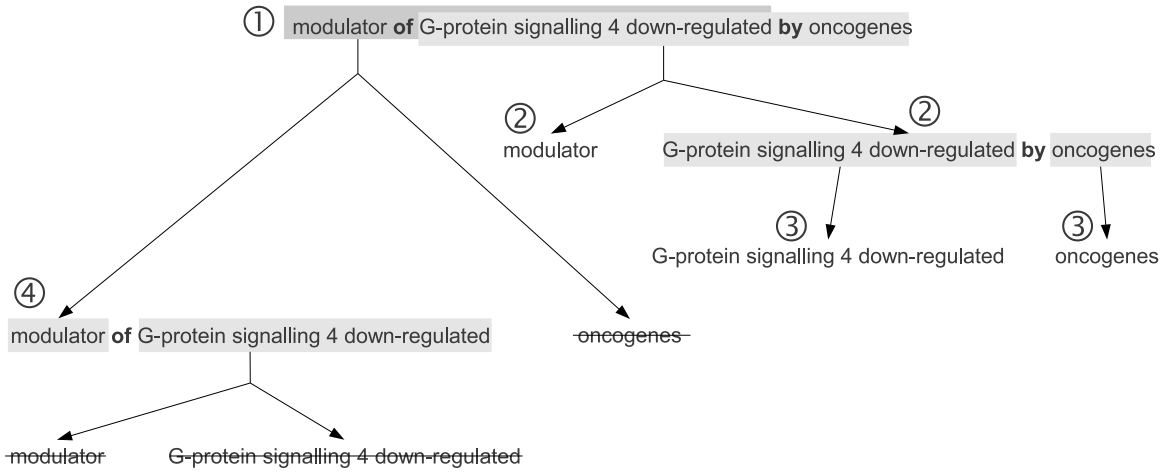
  A cette étape nous réutilisons les règles de transition de la grammaire d'EN afin de fractionner le syntagme d'entrée en sens inverse de construction. Les fragments du syntagme d'origine résultant de cette partition sont alors analysés de manière indépendante après délimitation des bornes de l'EN candidate grâce aux transformations que nous venons de décrire.

  Les fragments nouvellement générés ne sont pas toujours strictement équivalents à des blocs mais peuvent couvrir plusieurs blocs concomitants. Selon le niveau du syntagme d'entrée (3, 4 ou 5) et dans le cas où plusieurs séparateurs d'une même catégorie fonctionnelle (participe passé ou présent, préposition et conjonction de subordination ou conjonction de coordination respectivement) sont présents, nous générons les différentes combinaisons de blocs de part et d'autre de ces séparateurs. Les fragments prenant en compte le plus grand nombre de blocs sont analysés en premier. En cas d'égalité, le fragment en concurrence situé en fin du syntagme original est testé en priorité. Nous avons en effet noté que les ENs d'intérêt étaient localisées préférentiellement du côté droit des séparateurs. Cette observation n'est pas forcément consistante avec l'usage en langue anglaise de rejeter les termes les plus importants en début des constructions syntagmatiques.
  - Lorsqu'aucune EN n'a été identifiée au sein d'un fragment, ce dernier devient un syntagme qui sera à nouveau fragmenté et analysé en suivant la même procédure.
  - Lorsqu'une EN a été identifiée au sein d'un fragment, l'analyse récursive de ce dernier est arrêtée.

Il est à noter que des syntagmes préalablement extraits lors de combinaisons de découpage différentes ne sont pas évalués de nouveau.

La figure 3.7 illustre un exemple de processus de décomposition d'un syntagme à la recherche d'ENs. Dans cet exemple, le syntagme "G-protein signalling 4 down-regulated" n'est pas découpé grâce à une règle de transition $N_3 \rightarrow N_2$ avec comme verbe séparateur "signalling" car "4 down-regulated" n'est pas un bloc cœur.

FIG. 3.7 – Exemple de fractionnement des blocs d'ENs candidates



Les numéros associés à chaque syntagme indiquent leur ordre de génération. Les syntagmes rayés ne sont pas évalués car ils ont déjà été extraits.

3.2.3 Disambigüisation des ENs

Deux types d'ambigüités doivent être distinguées : celles liées aux problèmes de découverte des ENs dans le texte, qui sont résolues avant ou pendant la phase d'EEN, et celles en relation avec les questions d'identités alternatives des ENs, prises en considération à l'étape d'IEN. Nous allons décrire les méthodes mises en œuvre afin de répondre à ces deux difficultés.

Ambigüités liées à l'EEN. Les auteurs d'articles scientifiques peuvent à l'occasion redéfinir leurs propres alias alternatifs pour les ENs longues ou complexes. En général, cet alias original est précisé dès la première occurrence de l'EN dans le texte. Certains sont uniques et spécifiques à l'article et donc ne sont pas présents au sein de nos dictionnaires.

Les syntagmes basés sur un groupe nominal, entre parenthèses et précédés par une EN reconnue sont automatiquement associés en tant qu'alias de ce dernier. A cet effet, nous gardons pour chaque article un dictionnaire dynamique qui contient les alias découverts au gré de l'exploration du texte. Ce dictionnaire est particulier dans le sens où sa durée de vie est limitée à l'article en cours et est complémentaire des dictionnaires dits statiques, décrits dans la section 3.1, car il ne contient que des références de synonymie vers des entrées de ces derniers. Lors de l'étape d'IEN, les ENs candidates sont d'abord mis en correspondance avec les entrées de ce dictionnaire puis en cas d'échec avec celles des dictionnaires statiques. Un autre exemple d'ambigüité, à la fois en relation avec le processus d'EEN et d'IEN, est le problème des anaphores et co-références. Cette difficulté spécifique sera envisagée à part dans la section 3.2.5.

Ambigüités liées à l'IEN. Lorsqu'une EN a été identifiée grâce à nos dictionnaires, il nous est encore nécessaire d'évaluer la nature réelle d'une EN identifiée et ce pour différentes raisons :

- Tout d'abord, une entité peut être associée à différentes classes au sein de nos dictionnaires. Ce type d'ambigüité est une des formes d'homonymie décrites dans la section 2.4.2.3.
- Ensuite, un terme de l'anglais standard peut être incorrectement assimilé à un objet biologique d'intérêt. Par exemple, est-ce que le terme "Aim" est le nom du gène "absent in melanoma" ou un synonyme de but/objectif dans le texte ? Cette difficulté est détaillée dans la section 2.4.2.3.
- Finalement, deux entités distinctes de nos dictionnaires peuvent partager le même nom et être associées à la même classe. Par exemple, "CARP" peut soit désigner l'entité "carbonic anhydrase VIII" ou l"'ankyrin repeat domain 1 gene". Nous ne cherchons pas à résoudre ce type d'ambigüité dans notre étude à cause de sa relative rareté dans les nomenclatures humaines que nous manipulons. Parmi une sélection aléatoire de 1000 variantes orthographiques à partir de nos dictionnaires, nous n'avons décelé qu'un seul couple d'objets biologiques sans rapport et associé à une même entrée.

3.2.3.1 Désambigüisation de la classe d'une EN identifiée

L'information contextuelle, sous la forme des termes retrouvés dans l'environnement de l'EN, est utilisé afin de s'assurer de l'identité réelle de l'EN détectée. Nous avons assemblé

une liste de termes (de un à plusieurs mots), annotés à la main, et associés à des concepts utilisés dans le cadre de la biologie et plus particulièrement de l'expression de gènes. La présence de ces termes est significative d'un contexte soit expérimental (c'est à dire en référence à un protocole d'étude en biologie), physiologique ou pathologique particulier. Les concepts peuvent décrire une classe biologique explicitement (par exemple, 'gene', 'promoter', 'transcription factor') ou implicitement (par exemple, 'over-expression' pour un gène, 'transactivation' pour un facteur de transcription, 'secretion' pour une protéine ou un tissu ou organe). Chacun de ces termes peut être associé avec une ou plusieurs classes d'entité selon la précision de son contenu sémantique. La construction du lexique est détaillée ci-après.

Critères de désambigüisation. Nous utilisons d'abord une procédure à base de règles avec des contraintes grammaticales, et par extension syntaxiques, afin d'identifier les termes dans notre liste qui sont connectés à l'entité à désambigüiser. Selon les caractéristiques grammaticales et syntaxiques des termes connectés, un score est alors donné à chaque catégorie et une EN est désignée comme faisant vraisemblablement partie de la classe ayant le meilleur score. Le score relatif à une catégorie particulière est incrémenté à chaque fois qu'une occurrence d'un terme connecté à l'EN à désambigüiser et appartenant à cette classe spécifique est détectée (après normalisation). Les fonctions grammaticales permettant de définir quels sont les termes connectés à une EN sont de trois types :

- les compléments du nom (*noun adjuncts*) et différentes appositions mais aussi les adjectifs (*relational adjectives* et participes présents ou passés). Par exemple, "IL2 gene", "expression of IL2" ou "secreted IL2".
- Les verbes dont le sujet ou l'objet est l'EN. Par exemple, "abnormal cells overexpress IL2".
- La locution prépositive, avec participe, infinitive ou avec gérondif. Par exemple "Using ELISA-immunoassays, we detected an IL2 complex".

Les termes appartenant aux deux dernières catégories grammaticales sont découverts une fois que la syntaxe de la phrase, même partielle, a été résolue. Les méthodes relatives à l'exploration de la syntaxe sont décrites dans la section 4.1. Dans un souci de simplicité, nous ne définissons aucune relation d'ordre ou de priorité parmi les caractéristiques grammaticales sélectionnées dans la détermination de la classe de l'EN. Ainsi, les *POS* des termes de désambigüisation contextuelle ont le même pouvoir de discrimination. D'autre part, la position relative de ces termes par rapport à l'EN (par exemple, le terme est-il situé à gauche ou à droite de l'EN ? A quelle distance de l'EN le terme est-il positionné ?) n'a pas été retenue comme un critère d'aide à la désambigüisation.

Compatibilité de la classe contextuelle et de la classe vraisemblable. Les entités enregistrées dans nos dictionnaires sont toutes associées préalablement avec une ou plusieurs classes par défaut, définies en fonction de la ressource d'origine. Avant d'associer l'EN identifiée avec la meilleurs classe décrite par le contexte, nous cherchons à valider la vraisemblance de cette désambigüisation vis-à-vis des classes par défaut des dictionnaires. Par exemple, une entité que nous avons répertorié dans nos dictionnaires à partir d'un ressource terminologique spécifique aux gènes et aux protéines ne peut être détectée comme étant un tissu d'après le contexte. Les règles suivantes, reprises dans la figure 3.8, nous permettent de nous nous assurer de la compatibilité de la classe désambigüisée grâce au contexte avec les classes par défaut des dictionnaires :

- Si seule une classe est commune entre l'ensemble des classes détectées à partir du contexte et celles présentes dans les dictionnaires, alors l'entité est désambigüisée comme étant de cette classe particulière.
- Si plus d'une classe est commune entre le contexte et les dictionnaires, l'entité est désambigüisée comme faisant partie de la classe avec le meilleur score.
- En cas d'égalité ou si le contexte est nul, l'entité est considérée comme non-résolue et aucune classe ne lui est associée.
- Une exception notable à cette dernière règle est néanmoins admise : si une ambigüité demeure uniquement dans le choix de la classe 'gène' et 'protéine' ou de la classe 'site de liaison aux facteurs de transcription' et 'facteur de transcription', alors l'EN est désignée comme appartenant à la classe 'protéine' ou 'facteur de transcription' respectivement. Par expérience, nous avons constaté que les auteurs d'articles se réfèrent préférentiellement à la forme protéique lorsqu'aucune information contextuelle n'est donnée.
- Si aucune classe n'est commune entre le contexte et les dictionnaires, nous considérons que l'EN à été identifiée par erreur et est en conséquence étiquetée en tant que faux positif. Par exemple, dans la phrase "The cells have been incubated at **T** 4 degrees overnight.", une EN présentée en gras a été *a priori* identifiée en tant que 'T cell' et sa classe est spécifiée dans les dictionnaires comme étant 'cellule'. Or le contexte de l'EN, représenté par les termes soulignés, stipule que l'EN est de type 'autre' (c'est une classe particulière qui indique que ce n'est pas une EN de nos dictionnaires). L'EN est donc rejetée car la classe du contexte et des dictionnaires est différente.

Construction du lexique de désambigüisation. Le lexique de désambigüisation contient plus de 3800 entrées. Chaque entrée correspond à un terme, son rôle grammatical envisagé (nom ou adjectif, verbe et utilisé dans une locution) et la ou les classes de désambigüisation

associées. Le lexique a été élaboré semi-automatiquement à partir d'un sous-ensemble du corpus **GENIA** dans lequel nous avons préalablement ré-associé les ENs étiquetées avec les classes présentes dans nos dictionnaires. Nous avons extrait automatiquement de ce dernier un ensemble de termes en fonction de leur rôle grammatical vis-à-vis des ENs et ce grâce aux critères et aux méthodes définis ci-dessus. Ces termes ont ensuite été classés en fonction de leur fréquence de co-occurrence avec une classe biologique particulière. Finalement, nous avons expertisé et gardé dans notre lexique uniquement les associations les plus pertinentes et discriminantes, soit environ 3400 termes. Ce lexique a été complété de manière entièrement automatique avec 400 expériences et protocoles expérimentaux présents dans le **Metathesaurus UMLS** et par extension déjà répertoriés dans nos dictionnaires. Les techniques et protocoles expérimentaux sont spécifiques d'une ou d'un nombre restreint de classes biologiques de part leur nature. Par exemple, la "Polymerase Chain reaction" est une technique expérimentale utilisée uniquement dans le cadre de la caractérisation de gènes ou de petites portions d'ADN, en aucun cas elle ne permet de manipuler des protéines. Chaque description textuelle présente dans le **Metathesaurus UMLS** nous permet d'associer une classe biologique de désambigüisation à un nom de protocole expérimental. Nous utilisons pour cela le principe de co-occurrence des termes de la liste de désambigüisation déjà établie au sein de chaque description d'expérience : la classe biologique la plus représentée dans une description du **Metathesaurus UMLS** est systématiquement assignée à la classe de désambigüisation de l'expérience décrite. Les noms de protocoles sont considérés comme pouvant à la fois assumer le rôle d'un complément du nom ou être présent dans une locution. Une petite portion du lexique de désambiguïsation est proposée dans la table 3.6 en guise d'illustration.

Quelques raffinements ont été apportés au contenu de ce lexique :

- il est à noter d'une part que nous avons ajouter dans ce lexique certains termes particuliers, qui ne se réfèrent à aucune classe présente dans les dictionnaires. Ces termes sont associés à une classe spéciale et sont utilisés afin d'aider à filtrer les ENs faux positifs. Par exemple, les termes "viral" ou "bacterial" en tant que compléments du nom ou adjectifs peuvent préciser l'origine non-humaine de l'EN connectée. Si cette classe spéciale est prédominante dans le contexte, l'EN est considérée comme ayant été identifiée par erreur.
- D'autre part, nous avons dans notre lexique des termes 'dynamiques'. Ces termes ne sont pas des mots à proprement parlé mais des patrons d'expressions régulières. Par exemple, la présence d'un motif ^[ATCGUN]{4,}$, représentant une séquence nucléique, en tant que complément du nom peut être significatif d'un contexte lcoal

de type 'gène' ou 'site de liaison aux facteurs de transcription'.

3.2.3.2 Entités des dictionnaires et mots de l'anglais courant

Afin de distinguer les entités d'intérêt de leurs homonymes de l'anglais courant, nous examinons chaque entrée de nos dictionnaires en les comparant au contenu d'un dictionnaire anglais généraliste. Toute EN est considérée comme une source potentielle d'ambigüité avec l'anglais standard si chaque mot qui le compose est retrouvé dans un dictionnaire dérivé du **Webster 1913**. A l'étape de création de nos propres dictionnaires, chaque variante est marquée comme étant de nature ambigüe ou non grâce au contenu d'un dictionnaire [6] qui contient 91840 mots de la langue anglaise et non liés au domaine de la biologie. Les règles spécifiques à cette forme de désambigüisation, et illustrées dans la figure 3.8, sont détaillées ci-dessous :

- A l'étape de désambigüisation des classes des ENs identifiées, une EN dont la classe n'a pu être précédemment résolue, et qui n'est pas uniquement composée de mots de l'anglais standard, est alors assignée à la ou les classes associées à la dernière occurrence désambigüisée de cette même entité au sein du document en cours. Par expérience, nous avons observé une amélioration globale de 5% de la précision lorsque l'on suggère automatiquement une classe par défaut lorsque le contexte est trop pauvre pour prendre une décision. En revanche, ce gain moyen en précision disparait si l'on ne prend pas en compte la résolution des homonymes de l'anglais courant.
- Si aucune classe ou ensemble de classes n'a pu être attribué à une EN identifiée et si celle ci est marquée comme ambigüe dans nos dictionnaires, nous considérons que cette EN est un faux positif.
- Finalement, si aucune classe ou ensemble de classes n'a pu être attribué à une EN identifiée, si l'EN n'est pas marquée comme ambigüe dans nos dictionnaires et si uniquement une seule classe lui est assignée par défaut dans nos dictionnaires, l'EN est désambigüisée comme faisant partie de la classe par défaut proposée par les dictionnaires. Dans ce cas de figure particulier la classe de l'EN peut être résolue sans ambigüité malgré la présence d'un contexte local extrêmement pauvre.

3.2.4 Evaluation du système de REN

Nous avons mesuré les performances du système de REN de deux manières différentes, les deux protocoles utilisés ne répondent pas aux mêmes contraintes et aux mêmes ques-

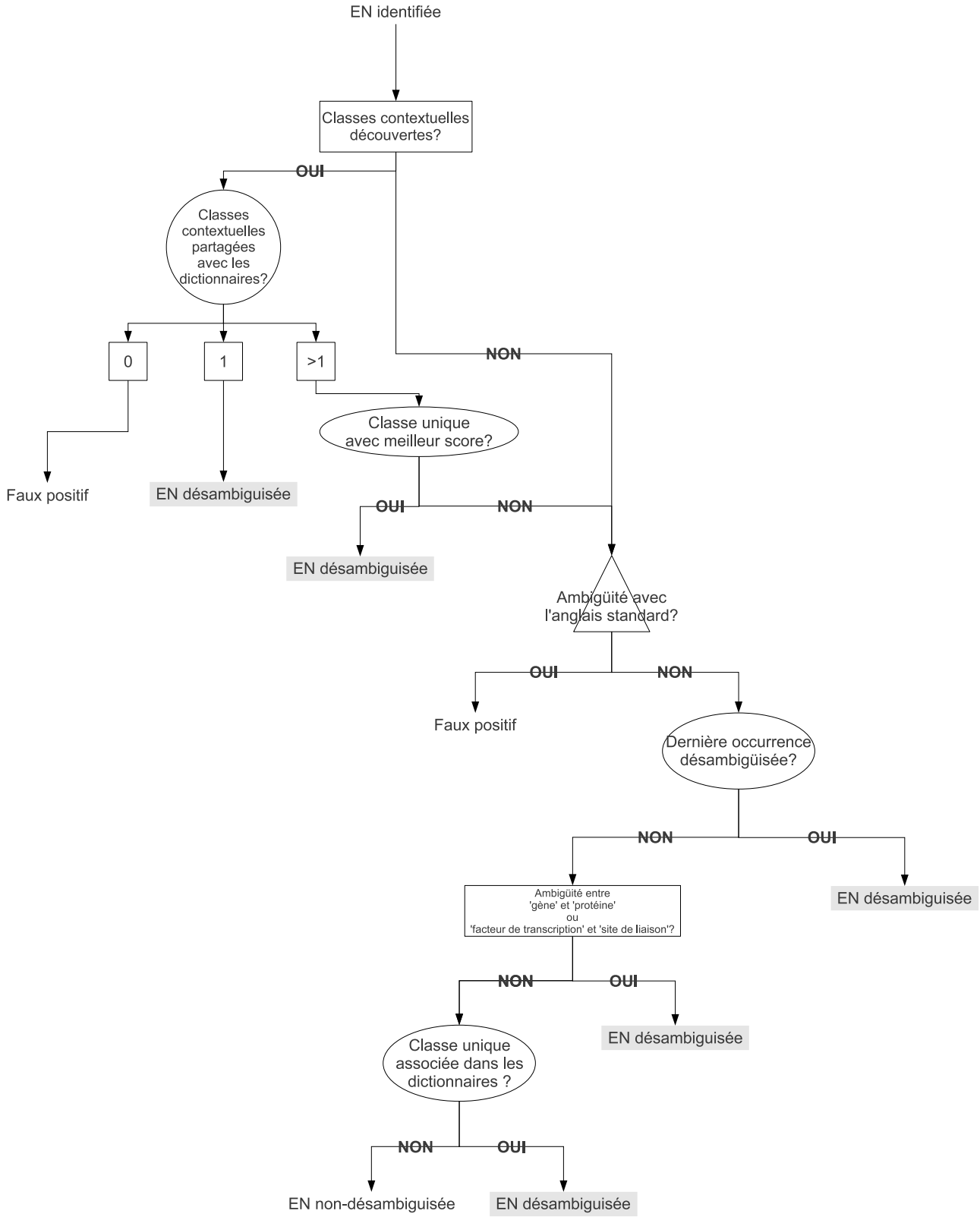
[6] http //humanities.uchicago.edu/orgs/ARTFL/forms_unrest/webster.form.html

TAB. 3.6 – Exemples d'entrées du lexique de disambigüisation des classes des ENs identifiées.

Terme	Fonctions grammaticales compatibles	Classes de désambigüisation correspondantes
3' flanking	CdN	Site de liaison aux facteurs de transcription, Gène
DNA binding	CdN	Facteur de transcription, Protéine
adherence	CdN	Cellule, Tissu ou organe
adrenal	CdN	Tissu ou organe
agonist	CdN	Facteur de transcription, Protéine
allele	CdN	Gène
antibody	CdN	Protéine
antiporter	CdN	Protéine
antisense	CdN	Gène
apoptosis	CdN	Tissu ou organe
atherosclerotic	CdN	Tissu ou organe
autoantibody-producing	CdN	protein
autosome	CdN	Gène
base pair	CdN	Site de liaison aux facteurs de transcription, Gène
binucleotide	CdN	Site de liaison aux facteurs de transcription
blast	CdN	Cellule
c-terminal	CdN	Facteur de transcription, Protéine
cDNA	CdN	Gène
cancerous	CdN	Cellule, Tissu ou organe
cassette	CdN	Site de liaison aux facteurs de transcription, Gène
cell	CdN	Cellule
centromere	CdN	Gène
cis activation	CdN	Site de liaison aux facteurs de transcription
clone	CdN	Cellule, Gène
co-culture	CdN	Cellule, Tissu ou organe
co-transporter	CdN	Protéine
coimmunoprecipitate	Vi	Facteur de transcription, Protéine
consensus	CdN	Site de liaison aux facteurs de transcription
copy	CdN	Gène
decidualized	CdN	Cellule, Tissu ou organe
degranulate	Vd	Cellule
desphosphorylation	CdN	Facteur de transcription, Protéine
differentiation	CdN	Cellule
dominant negative	CdN	Gène
downregulate	Vi	Gène
downstream	CdN	Site de liaison aux facteurs de transcription
encode	Vi	Facteur de transcription, Protéine
enhancer	CdN	Site de liaison aux facteurs de transcription
exon	CdN	Gène
gene encoding	CdN	Facteur de transcription, Protéine
homodimerize	Vd	Facteur de transcription
immortalize	Vi	Cellule
lineage	CdN	Cellule
mAb	CdN	protein
population	CdN	Cellule

CdN = complément du nom, adjectif ou apposition de l'EN, Vd = verbe dont le sujet est l'EN, Vi = verbe dont l'objet est l'EN, Loc = locution connectée avec l'EN.

FIG. 3.8 – Organigramme du processus de désambiguisation



tions. Dans une première partie nous présentons les résultats de nos méthodes d'EEN que nous comparons à d'autres systèmes sur un corpus de test générique puis, dans une deuxième partie, nous mesurons les performances combinées des processus d'EEN et d'IEN sur un jeu de données adapté à la problématique de la régulation de gènes. Il est important de noter que nous ne cherchons pas à résoudre les anaphores et co-références.

3.2.4.1 EEN

Nous avons tout d'abord évalué le système grâce au protocole **BioNLP/NLPBA 2004**[7]. Seules les étapes d'EEN et d'association à une classe biologique sont ici mesurées. La mise en correspondance des ENs extraites avec les entités de nos dictionnaires n'est pas évaluée par le biais de ce protocole. **BioNLP/NLPBA 2004** est une compétition internationale sur le thème de la REN dans le domaine de la biologie moléculaire et plus précisément de la classification des termes techniques correspondant à des concepts retrouvés dans **Medline**. Les données d'apprentissage et d'évaluation utilisées pour cette tâche proviennent du corpus **GENIA 3.02**. Lors d'une étude préliminaire du comportement de notre système vis-à-vis des données fournies et au vu des résultats escomptés, nous avons constaté certaines limitations inhérentes à nos méthodes les rendant non parfaitement conformes aux attentes du protocole. Il est à noter que nos méthodes d'EEN et d'IEN sont étroitement inter-dépendantes. L'échec de l'étape d'IEN à pour conséquence que l'entité n'est pas extraite des textes, et ce même si l'EN est valide mais uniquement absente de nos dictionnaires. Une autre considération importante qui nous place en défaut vis-à-vis de la mise en conformité au protocole **BioNLP/NLPBA 2004** est que nous n'utilisons pas les données d'apprentissage fournies, uniquement les données d'évaluation. D'autres différences ponctuelles doivent être prises en considération et harmonisées, cette fois plus précisément vis-à-vis du corpus de test :

- D'une part, lors du développement des méthodes, nous n'étions uniquement intéressés que par la détection des individus qui constituent un classe d'objet biologique particulier dans les textes. Or conceptuellement **GENIA** ne définit pas les classes de la même manière. En effet, les classes **GENIA** ne distinguent pas les noms des individus des noms des familles ou des groupes auxquels ils appartiennent. De la même façon, des portions et sous-unités d'un individu sont aussi regroupées au sein de la même classe. Par exemple, "hormone binding domain" a été étiqueté comme une 'protéine' dans les données de test alors que nous considérons l'expression trop

[7] http //www-tsujii.is.s.u-tokyo.ac.jp/GENIA/ERtask/report.html

imprécise pour être acceptée telle quelle. Ainsi, nous avons adapté spécifiquement nos méthodes pour cette tâche. Nous considérons que les syntagmes qui contiennent des termes présents dans notre lexique de désambigüisation sont des instances valides d'une classe. De plus, nous avons dupliqué dans le cadre de cette évaluation les formes variantes de nos dictionnaires se terminant par un numérique, un caractère alphabétique unique ou un symbole que nous avons supprimés lors de la copie. En effet, ces marqueurs sont les identifiants spécifiques des membres d'un même groupe, nous partons de l'hypothèse que leur suppression permet de ne garder que le nom de la famille ou du groupe d'appartenance dans la forme variante. Par exemple, la variante "Interleukin 2" originellement présente dans nos dictionnaires va permettre de générer une nouvelle variante : "Interleukin" qui est sensée représenter la famille d'appartenance de l"'Interleukin 2". La combinaison de ces deux techniques, très simples, améliore néanmoins le *score F* de près de 20% globalement et permet de se placer dans des paramètres acceptables vis-à-vis de la mise en conformité au protocole **BioNLP/NLPBA 2004**.

- D'autre part, pour la tâche imposée par **BioNLP/NLPBA 2004**, seules 5 classes biologiques parmi les 36 proposées par **GENIA** ont été retenues. Malheureusement, certains compromis ont été nécessaires concernant les classes biologiques proposées d'une part par **GENIA** et d'autre part par nos dictionnaires. Les catégories génériques **GENIA** 'protein' et 'DNA' ont été respectivement "mappées" avec nos propres classes 'protéine' + 'facteur de transcription' et 'gène' + 'site de liaison aux facteurs de transcription' dont les définitions sont plus restreintes. Inversement, les catégories **GENIA** 'cell line' et 'cell type' ont été fusionnées en une catégorie unique 'cell' afin de correspondre à notre propre classe 'cellule'. Nous n'avons pas non plus cherché à étendre nos dictionnaires afin de prendre en compte la classe **GENIA** 'RNA'. Comme précisé dans la section 3.1.3, le nettoyage du contenu d'un dictionnaire est un tâche lourde extrêmement couteuse en temps et en ressource humaine.

Les résultats sont montrés dans le tableau 3.7. Seuls les scores correspondant à la sous-tâche *complete match* y sont figurées. Les bornes de chaque EN extraite du texte doivent être strictement identiques à celles présentes dans les données d'évalutation. Le calcul des métriques est effectué automatiquement via l'outil d'évaluation fourni par les organisateurs de la compétition. Nous avons présenté dans le tableau les résultats obtenus par deux compétiteurs parmi les 10 meilleurs retenus par **BioNLP/NLPBA 2004**. Zho04 [ZS04] a été classé premier de la compétition. Les auteurs utilisent un système d'apprentissage automatique basé sur les *HMM*. Les caractéristiques retenues pour l'apprentissage

sont : les *têtes* des groupes nominaux, les *POS*, les différents affixes du groupe nominal (correspondant grossièrement à nos préfixes et à nos suffixes) et la différentiation majuscule/minuscule et la présence de symboles numériques. En complément, un dictionnaire établi à partir des données d'entrainement et des bases de données **SwissProt** et **LocusLink** sert à retrouver les alias d'ENs potentielles. La présence ou l'absence de l'expression au sein du dictionnaire est utilisée en tant que caractéristique supplémentaire dans le modèle *HMM*. En revanche, les auteurs se reposent entièrement sur des règles syntaxiques expertisées afin de découvrir les entités *en cascade* et de résoudre les phénomènes d'acronymie et d'abréviation. Lee04 [LHC04] a été classé dixième et utilise un système d'apprentissage basé sur les *SVM*. Les caractéristiques de leur modèle sont : la présence ou non d'un mot dans un nom de gène/protéine/espèce/tissu préalablement détecté (cette donnée est fournie par l'étiqueteur d'entités biologiques tierce **Yapex** [KGF+02]), la présence de symboles ou de ponctuations, le *POS* et la distance et fréquence du mot par rapport à un terme spécifique d'une classe biologique et référencé dans un lexique créé spécifiquement pour la tâche. Il est à noter que tous les systèmes représentés dans la compétition sont à base d'algorithmes d'apprentissage automatique. En se basant sur le *score F* obtenu, nous nous positionnons en sixième position du classement, et ce malgré l'absence de techniques d'apprentissage automatique dans notre approche.

TAB. 3.7 – Performance du système d'EEN et d'identification des classes biologiques à partir des données BioNLP/NLPBA 2004

Auteur/Méthode	Classe	Rappel	Précision	Score F
Zho04 [ZS04]	protein	79.24%	69.01%	73.77%
	DNA	73.11%	66.84%	69.83%
	cell	79.98%	79.06%	79.52%
Notre système	protein	53.03%	76.35%	62.58%
	DNA	32.84%	64.21%	43.45%
	cell	67.56%	81.02%	73.68%
Lee04 [LHC04]	protein	62.13%	46.14%	52.96%
	DNA	28.88%	44.79%	35.12%
	cell	64.77%	46.55%	54.17%

3.2.4.2 IEN

Dans une deuxième partie de l'étude, nous avons mesuré à la fois les performances des processus d'EEN et d'IEN de notre système. Nous avons sélectionné aléatoirement 100 résumés parmi les 404 proposés dans le corpus **GENIA** utilisé par **BioNLP/NLPBA**

2004 avec la contrainte supplémentaire qu'ils doivent inclure le mot clé **MeSH** "cytokine". Ces résumés ont alors été ré-annotés dans le but de correspondre uniquement aux classes répertoriées dans nos dictionnaires. D'autre part, les noms de famille, de groupe, de sous-structure, de sous-unité ou de complexe biologique n'ont pas été conservés. Nous nous sommes aussi efforcés de faire disparaitre de l'annotation les entités d'origine non-humaine. Chaque instance des ENs annotées a été spécialement marquée afin de rendre compte de l'identifiant qui lui est associé dans nos dictionnaires s'il est présent. Nous considérons en tant que vrai positif les ENs dont l'identifiant de dictionnaire et la classe associée sont strictement équivalents à ceux annotés dans les données d'évaluation. Si une EN annotée n'est pas extraite, identifiée ou désambigüisée, alors l'EN est étiquetée en tant que faux négatif. Une EN non-annotée qui a été néanmoins extraite et identifiée par la système est reconnue comme étant un faux positif. Une EN, incorrectement désambigüisée (par exemple annotée en tant que 'gène' mais associée par le système à la classe 'protéine') ou connectée à un identifiant de dictionnaire erroné (nous n'avons pas pénalisé les homonymies intra-classes qui ne sont pas gérées par le système, uniquement les homonymies inter-classes : c'est à dire deux entités qui partagent le même nom mais qui ne possèdent aucune classe pré-définie dans les dictionnaires en commun), est à l'opposée considérée en tant que faux négatif et faux positif.

Voir le tableau 3.8 pour les résultats. Dans cette étude, les entités correctement identifiées mais appartenant aux mammifères et non juste à l'humain sont néanmoins considérées comme des vrais positifs. Lors de la ré-annotation nous n'avons pas pu nous assurer de l'origine de toutes les ENs présentes.

TAB. 3.8 – Performance du système d'EEN et d'IEN à partir d'un sous-ensemble enrichi du corpus GENIA

	G	P	F	S	C	O	E	TOTAL
Précision	0.78	0.74	0.72	0.65	0.93	0.87	0.85	0.77
Rappel	0.75	0.71	0.68	0.63	0.71	0.87	0.71	0.71
Nombre de vrais positifs (VP)	476	630	561	123	530	7	92	2419
Nombre de faux positifs (FP)	36	74	72	27	16	1	10	236
Nombre de faux négatifs (FN)	64	103	109	32	189	1	31	529
Nombre d'ENs à la fois FP et FN	92	146	144	39	23	0	5	449

G = gènes, P = protéines (sans les facteurs de transcription), F = facteurs de transcription, S = sites de liaison aux facteurs de transcription, C = types et lignées cellulaires, O = tissus et organes, E = protocoles et techniques expérimentaux

3.2.4.3 Analyse des erreurs

Lors de l'analyse des principales sources d'erreur en rapport avec la tâche **BioNLP/NLPBA 2004**, nous avons observé que la plupart des difficultés rencontrées étaient dues aux disparités entre les méthodes d'étiquetage de **GENIA** et la structuration de nos méthodes d'EEN. Tout d'abord, certaines ENs sont parfois imbriquées au sein d'autres ENs plus larges (*en cascade*) et ont été annotées séparément. Par exemple, l'expression complète "NF Kappa B binding sites" est correctement assignée à la classe 'DNA' alors que la portion "NF Kappa B" incluse est parfois, mais non systématiquement, marquée comme étant 'protein'. Ensuite, la présence de termes satellites placés à gauche de l'EN (préfixes) et qui peuvent ou non faire partie du nom de l'entité selon l'annotateur est une difficulté non négligeable. Par exemple, l'expression "affinity-enriched NF A2" est parfois interprétée, au gré des documents, comme une EN autonome, d'autre fois seule la portion "NF A2" est considérée comme étant l'EN. Une autre source importante d'erreurs est due à la présence de noms de familles, groupes, sous-structures, sous-unités et complexes dans l'annotation. Notre système ne gère pas correctement de tels objets biologiques et notamment les bornes de ces ENs. En conséquence, nous nous sommes libérés de ces limitations lors de la ré-annotation d'un sous-ensemble du corpus **GENIA** dans le but de mesurer les performances réelles de notre système de REN. Durant l'analyse des des performances globales du système (EEN + IEN), nous avons constaté que les erreurs pouvaient être aisément classées en un nombre restreint de catégories.

La principale source de faux positifs est à la fois liée à des problèmes de désambigüisation et à des contradictions au sein des dictionnaires. Concernant les problèmes spécifiques aux dictionnaires, il n'est pas rare que des noms génériques y aient été consignés par erreur. Par exemple, "immediate early gene" est le nom d'une famille de gènes et est néanmoins présent dans nos dictionnaires. Les ENs faussement identifiées correspondent en très large majorité à des portions de l'expression représentant l'entité réelle. Par exemple, "pseudo tumor necrosis factor receptor" est absent des dictionnaires et a été associé par défaut à "tumor necrosis factor receptor" alors que les deux entités sont distinctes. La principale source de faux négatifs est due à l'absence de l'EN au sein des dictionnaires (ceci est plus particulièrement observé auprès des dictionnaires construits à partir du **Metathesaurus UMLS**). Deux autres causes importantes de faux négatifs sont, d'une part l'absence des formes variantes correspondant aux ENs extraites des textes au sein de nos dictionnaires, et d'autre part la pauvreté du contexte local de l'EN (tel que défini dans le paragraphe 3.2.3 et utile à la disambigüisation des classes des ENs). Il est important de souligner que

la très grande majorité des erreurs comptabilisées à la fois en tant que faux positifs et faux négatifs sont les conséquences d'une désambigüisation incorrecte. La difficulté majeure observée est l'absence des termes clefs à la disambiguïsation au sein du lexique lorsque le contexte est riche d'autre termes reconnus mais de sémantique différente. Par exemple, dans l'expression "gene activity of *NF-Kappa-B*", "NF-Kappa-B" est désambigüisé en tant que gène, or l'entité est ici décrite sous la forme protéine. Le terme "gene activity" implique que l'EN n'est pas sous sa forme 'gène' même si le texte traite effectivement de l'activité du gène. Nous avons pu constater que l'absence d'ordre et priorité au sein de l'ensemble des termes de désambigüisation détectés (et notamment l'assignation de poids en fonction de la distance à l'EN) ne nous a pas globalement pénalisé. Ceci est majoritairement du à la complémentarité de deux facteurs : d'une part il est rare que le contexte local contiennent plus de trois termes de désambigüisation simultanément, et d'autre part les classes décrites par ces termes sont généralement consistantes entre elles. Il arrive aussi fréquemment que les variantes référencées dans nos dictionnaires appartiennent à plusieurs identifiants distincts dans les dictionnaires alors que ces derniers représentent finalement la même entité. L'expertise du contenu des dictionnaires est une tâche essentielle qui ne doit pas être négligée. En revanche, il nous a été permis d'observer que les étapes de normalisation et de génération automatique de variantes (voir le paragraphe 3.1.2.2) ne généraient pas d'erreurs sur le corpus de test. Aucune ambigüité additionnelle n'est apportée par ces techniques : les formes variantes générées sont uniques. Les résultats les plus faibles sont obtenus lors du traitement des sites de liaison aux facteurs de transcription. Deux difficultés majeures en sont responsables. D'une part, les noms des sites de liaison aux facteurs de transcription sont relativement courts et ambigus. Par exemple, "A Box" et "C" sont deux noms des sites de liaison aux facteurs de transcription. Leur nomenclature est la plus pauvre, la moins structurée et la moins supportée de toutes. D'autre part, les références aux sites de liaison aux facteurs de transcription dans les textes sont souvent implicites. Par exemple, les auteurs de publications utilisent fréquemment la séquence nucléique "TATAAA" pour symboliser l'entité "TATA Box". De telles séquences nucléiques sont extrêmement variables d'un article à l'autre de part leur nature biologique. La séquence nucléique, et donc la combinaison de caractères A, T, C et G qui la décrit, n'est pas identique d'un gène à l'autre.

Ces résultats sont prometteurs et ce malgré le biais évident introduit par ce corpus de test de 100 résumés. En effet, la représentation quantitative du nombre de classes est relativement disproportionnée, de plus le contenu du corpus se révèle homogène : les entités biologiques rencontrées dans ce sous-ensemble de documents sont très souvent

identiques.

3.2.5 Résolution des anaphores et de co-références

La résolution des co-références et des anaphores (voir le paragraphe 2.4.2.2) nous permet de ne plus nous limiter à l'IEN et de rendre possible l'identification des références implicites aux entités dans les textes. Sur un total de 868 phrases extraites d'un sous-ensemble du corpus de test utilisé pour l'identification des relations entre ENs (voir la section 4.4.1), nous avons détecté 18 co-références ou anaphores et 299 ENs soit un rapport approximatif d'une référence implicite à une EN pour 16 entités explicitement nommées.

Nous proposons une méthode de résolution des anaphores et des co-références basée sur les travaux de Lappin *et al.* [LL94] pour la langue anglaise. La méthode de Lappin *et al.* permet de résoudre les pronoms à la troisième personne (dans leur *cas* nominatif, par exemple "he", accusatif, par exemple "him" et possessif, par exemple "his"), certaines anaphores lexicales (les formes *réflexives*, par exemple "myself" et "yourself", et les formes *réciproques*, par exemple "each other" et "one another") ainsi que les pronoms *explétifs* (par exemple "it is raining" et "it has been demonstrated that(...)") dans les textes du domaine général. L'algorithme original ne prévoit pas de résoudre les co-références.

Nous utilisons l'implémentation proposée par **JavaRap**[8] [QKC04] que nous avons étendu pour nos besoins. L'algorithme nécessite que la syntaxe de la phrase, même partielle, ait été préalablement résolue. Les méthodes relatives à l'exploration de la syntaxe sont décrites dans la section 4.1. **JavaRap** utilise un système de pondération des solutions (appelées antécédents) candidates d'une anaphore dérivé de la structure syntaxique des phrases et un modèle dynamique basé sur les intentions du discours afin de sélectionner l'antécédent le plus vraisemblable parmi tous ceux disponibles. Aucune condition d'ordre sémantique n'est mise en œuvre. Les principaux composants de **JavaRap** sont :

- un filtre syntaxique à l'intérieur d'une même phrase afin de déterminer les dépendances anaphoriques d'un pronom sur des bases purement syntaxiques.
- Un filtre morphologique afin de déterminer les dépendances anaphoriques d'un pronom en fonction de sa compatibilité en genre (masculin, féminin, neutre), personne (je, tu, il, nous, etc) et nombre (un ou plusieurs).
- Une procédure pour identifier les pronoms explétifs.

[8] http //www.comp.nus.edu.sg/~qiul/NLPTools/JavaRAP.html

- Un algorithme pour retrouver l'antécédent le plus probable d'une anaphore lexicale à l'intérieur d'une même phrase.
- Une procédure afin d'assigner des pondérations aux antécédents candidats en fonction de leur rôle grammatical, du parallélisme des fonctions grammaticales, du nombre de fois qu'ils sont mentionnés et de leur proximité vis-à-vis de l'anaphore. Les poids les plus forts sont assignés en priorité :
  1. aux sujets des verbes,
  2. aux objets directs vis-à-vis des autres compléments,
  3. aux arguments d'un verbe (par exemple, "The protein activates RhoA. *It* was then phosphorylated.") par rapport aux compléments du verbe (par exemple, "The protein activates near the RhoA domain. *It* was then phosphorylated.") et aux objets de syntagmes prépositionnels (par exemple, "It indicates that RhoA is activated. *It* was then phosphorylated.").
- Une procédure pour identifier les groupes nominaux liés aux anaphores grâce à la somme des poids de chacun des éléments du groupe nominal.
- Un critère de décision dans le but de sélectionner l'antécédent candidat le plus vraisemblable parmi une liste.

L'ensemble de ces composant est décrit en détail dans l'article de présentation de **JavaRap** [QKC04].

Nous avons étendu l'algorithme original afin de prendre en compte certaines co-références et non juste les anaphores. Nous ne nous intéressons qu'aux co-références pouvant potentiellement se référer aux ENs d'intérêt du texte. Les co-références admissibles à la résolution sont des groupes nominaux qui ne sont pas des ENs et composés uniquement d'un déterminant (par exemple, "the", "these", "each of the"), d'un chiffre ou d'un adjectif facultatif (par exemple, "three", "corresponding") suivi d'un terme issu d'une liste comprenant les compléments du nom contenus dans notre lexique de désambigüisation (voir la section 3.2.3.1). Nous avons aussi raffiné la procédure d'attribution des poids aux groupes nominaux antécédents candidats en ajoutant deux nouveaux critères. Ces critères ne s'appliquent qu'aux co-références et non aux anaphores.

- *La plus longue séquence commune.* Le ou les antécédents candidats qui contiennent la plus longue séquence de mots en commun avec la co-référence sont privilégiés. Par exemple, "IL2 antigen" a un poids supérieur à "IL2 molecule" dans la résolution de la co-référence "the antigen".
- *La compatibilité des classes biologiques.* Ce filtre constitue l'unique apport sémantique à l'algorithme. Les termes issus du lexique de désambigüisation sont tous asso-

ciés à une ou plusieurs classes biologiques particulières. Nous considérons qu'une EN désambigüisée qui partage une classe en commun avec la co-référence doit être préférentiellement mise en avant en tant qu'antécédent vraisemblable. Par exemple, "IL2 gene product" a un poids supérieur à "IL2 locus" dans la résolution de la co-référence "the protein".

La capacité du système à résoudre les anaphores et co-références a été étudiée sur le corpus de test présenté en introduction de ce paragraphe. Nous avons obtenu une précision de 54%. L'implémentation originale de **JavaRap** atteint 45%, toutes les co-références du texte ont été comptabilisées en tant que faux négatifs parceque **JavaRap** ne prend en compte que les anaphores. La précision de **JavaRap** dans le domaine de l'anglais général a été estimée à 58% par ses auteurs. Au vu des résultats obtenus, nous n'avons pas souhaité intégré le module de résolution des anaphores dans notre système de REN.

3.2.6 Résumé

Nous avons présenté une méthode simple d'extraction et d'identification d'ENs complexes dans le cadre de la biologie moléculaire à l'aide de dictionnaires. La principale limitation des approches à base de dictionnaires est l'impossibilité de détecter des entités inconnues. En revanche, les dictionnaires sont nécessaires pour l'identification des objets manipulés dans les textes. Le principal avantage des techniques exposées dans cette section est la caractère relativement générique de ces dernières dans le domaine biomédical. Le lexique de désambigüisation est le seul composant spécifiquement élaboré en fonction des classes biologiques contenues dans nos dictionnaires. L'étude des résultats obtenus lors de l'IEN sur notre corpus de test nous permet de considérer avec confiance que nous nous situons globalement parmi les bonnes voire très bonnes méthodes de REN actuellement disponibles. Peu de systèmes permettent encore de détecter d'autres ENs que les protéines et les gènes. Nous sommes, à notre connaissance, les seuls à nous intéresser à la détection des facteurs de transcription et à leurs sites de liaison sur les zones promotrices de gènes. Cela a pour principal désavantage de ne pas pouvoir comparer objectivement les performances du système de REN avec des solution concurrentes. Nous avons constaté que les performances globales du système d'IEN étaient essentiellement liées à la qualité du contenu des dictionnaires utilisés. Nous n'avons pas recouru aux techniques d'apprentissage automatique car nous n'avions pas de corpus d'apprentissage à la fois conséquent et adapté aux classes biologiques que nous manipulions lors du développement des méthodes. En revanche, il est tout à fait envisageable d'adapter les heuristiques de désambigüisa-

tion en tant que caractéristiques d'un modèle d'apprentissage lorsque ce corpus sera mis à disposition. Cela permettrait assurément de passer outre les limitations actuelles de nos techniques de désambigüisation, encore perfectibles. Le problème de la résolution des anaphores et des co-références reste encore en suspens.

Chapitre 4

Identification des relations entre ENs

Dans cette section nous nous attachons à retrouver les relations qui animent les différentes ENs identifiées dans les textes.

Nous détaillons ici une procédure qui permet à la fois de récupérer les liens d'intérêt qui unissent les ENs détectées et de découvrir certaines caractéristiques associées à ces relations. Cette première tâche est similaire à la tâche **MUC's Template Relation**[1] alors que la deuxième est apparentée à la tâche **MUC's Template Element**[1]. Néanmoins, il est à noter que ces deux tâches partagent les mêmes méthodes au sein de notre système.

Afin d'extraire ces informations des textes, nous employons une approche hybride combinant traitements syntaxiques et sémantiques. Bien que l'injection de sémantique très tôt lors du traitement des phrases conduit d'ordinaire aux meilleurs résultats décrits dans la littérature [MCSM04], la principale conséquence liée à ce choix est la très grande spécialisation et donc la forte dépendance du système d'IE à un domaine donné [YMTT05]. Un tel système est alors généralement très difficilement adaptable à un autre domaine d'application. Nous avons décidé d'arranger un compromis entre performance et adaptabilité.

Afin d'extraire les relations qui unissent les ENs dans les textes, trois composantes majeures sont requises :

1. D'une part, une hiérarchie de concepts manipulés dans le textes doit être dressée.
2. D'autre part, la définition d'une structure organisant la connaissance liée à ces

[1] http ://www-nlpir.nist.gov/related_projects/muc/proceedings/ie_task.html

relations doit être élaborée.

3. Finalement, des règles de correspondance permettant de transformer les portions d'intérêt d'une phrase en éléments des structures de connaissance doivent être mises au point.

Nous montrons que l'acquisition de la sémantique à partir de la syntaxe peut être scindée en différentes phases afin réduire la complexité souvent associée avec la conception de règles d'extraction spécifiques au domaine d'étude.

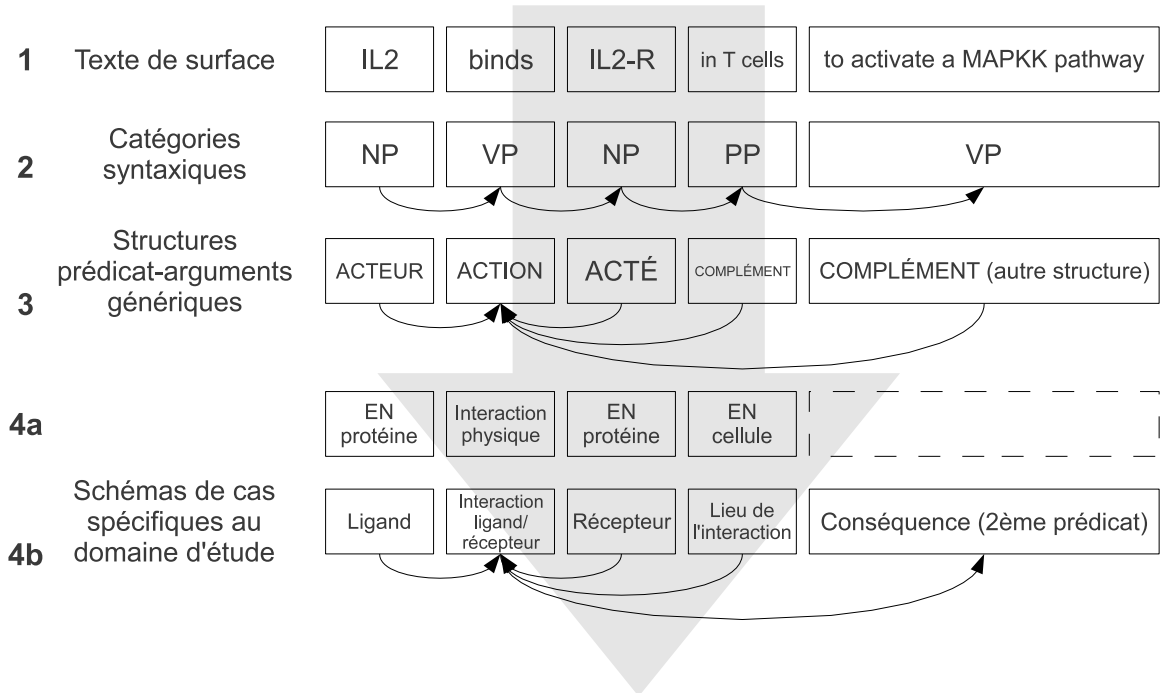


Fig. 4.1 – Les différents niveaux de représentation syntaxiques et sémantiques du texte de surface

La figure 4.1 illustre les différentes niveaux de représentation syntaxiques et sémantiques d'un texte. Nous analysons dans un premier temps la syntaxe des documents de manière fine [2]. L'ensemble des connaissances exploratoires auxquelles nous avons accès est modélisé grâce à des *structures prédicat-arguments* génériques (voir la section 2.1.4). Ces *structures prédicat-arguments* décrivent une première représentation sémantique, indépendante du domaine, de la phrase [3]. Finalement, nous utilisons des règles d'extraction spécifiques au domaine de la régulation de gènes à partir de ces *structures prédicat-arguments* afin de détecter des relations pré-catégorisées et biologiquement significatives entre les ENs identifiées [4a et 4b]. A ce stade, les informations d'intérêt alors extraites sont accessibles sous la forme de *schémas de cas* (voir la section 2.1.4) compa-

tibles avec notre modèle de la régulation de l'expression de gènes.

Dans la dernière partie de la section, la performance du système de recherche de relations entre ENs est évaluée en utilisant un corpus spécialisé et annoté à partir d'articles complets du domaine de la biologie moléculaire.

4.1 Acquisition de la syntaxe

Nous utilisons **Link Grammar Parser (LGP)** (voir l'annexe 4.5) afin de retrouver la structure syntaxique des phrases extraites des documents.

La syntaxe, une fois acquise par **LGP**, est représentée sous la forme d'une ou de plusieurs listes distinctes de *liens* établis entre paires de mots de la phrase. Pour référence, la liste des *liens* **LGP** est proposée dans la documentation en ligne de l'analyseur syntaxique[2]. La présence de conjonctions de coordination dans une phrase induit la création d'autant de listes autonomes de *liens* que d'éléments coordonnés. Chacun de ces ensembles de liens sera appelé *définition* de la phrase dans le reste du document. Par exemple, dans la phrase "IL2R can bind IL2, its natural ligand, **and** IL5", **LGP** va produire deux *définitions* de la syntaxe, une correspondant à la phrase "IL2R can bind IL2, its natural ligand" et l'autre à "IL2R can bind IL5". Ce que nous appellons *définition* dans ce document est l'équivalent de la *proposition* dans la logique des prédicats (voir la section suivante). A partir d'une *définition*, une représentation sous forme d'arbre peut être calculée et est appelée *arbre des constituants*, ou *arbre des syntagmes* dans ce présent document. Un ensemble de règles internes à **LGP** permettent de regrouper les mots de la phrase en fonction de leur propriétés fonctionnelles et des *liens* qu'ils établissent entre eux en syntagmes adjectivaux, nominaux, verbaux ou adverbiaux (les *constituants*). Les syntagmes sont ensuite arrangés hiérarchiquement dans un arbre selon un mode de représentation équivalent au *schéma* $\bar{X}$ (voir la section 4.5). Les noms (POS) des propositions ou syntagmes usuels utilisés dans le reste du document et compatibles avec ceux proposées par le projet **Penn Treebank** sont succinctement exposés dans l'annexe 7.

[2] http ://www.link.cs.cmu.edu/link/dict/summarize-links.html

4.1.1 Adaptation de LGP au domaine biomédical

Comme précédemment souligné dans la littérature [PSAN06], **LGP** présente des difficultés à analyser la syntaxe des textes techniques et notamment à cause du vocabulaire très spécialisé. **LGP** a été développé à l'origine pour analyser la syntaxe des transcriptions de dialogues téléphoniques en langue anglaise. Afin de créer des *liens* entre deux mots, **LGP** infère les fonctions syntaxiques possibles de ces mots dans la phrase à partir du vocabulaire présent dans un lexique contrôlé. **LGP** utilise trois méthodes, en cascade, pour gérer le vocabulaire des phrases : la recherche dans un dictionnaire, l'utilisation d'une classe de mot générique correspondant à une fonction grammaticale particulière à partir de certaines caractéristiques morphologiques remarquables (par exemple, les mots se terminant par '-ly' sont associés à une classe de mots générique qui rassemble les adverbes), l'utilisation en dernier ressort d'une classe de mot 'fourre-tout' lorsque le mot est inconnu. Cette dernière classe de mots définit toutes les combinaisons de *liens* possibles avec un verbe, un nom et un adjectif. Cette approche permet en théorie de toujours générer la combinaison correcte de *liens* pour des mots non présents dans le dictionnaire de **LGP**. En revanche, sur les textes techniques, le nombre de mots inconnus dans une phrase est très grand et conduit à l'explosion combinatoire de nombre de *liens* à tester afin de retrouver la syntaxe sous-jacente de la phrase. Les performances de **LGP** s'en trouvent diminuées. Il est aussi à noter que l'analyse d'une phrase donnée par **LGP** est limitée dans le temps : si une phrase n'a pu être traitée dans un intervalle de temps donné, l'analyse est alors effectuée rapidement en utilisant des paramètres extrêmement restrictifs mais qui conduisent inéluctablement à une solution approximative.

Afin de répondre à la difficulté que présente l'inadaptation du lexique de **LGP** au domaine biomédical, nous avons décidé de coupler l'utilisation de l'étiqueteur spécialisé **GENIA Part Of Speech Tagger** à **LGP**. Le *POS* retrouvé par **GENIA Part Of Speech Tagger** est utilisé lorsqu'un mot est absent du lexique **LGP** et permet de fournir les conditions adéquates d'établissement d'un *lien* pour l'analyseur syntaxique. Chaque *POS* est préalablement associé à une entrée dans le lexique **LGP**, la définition correspondante se rapporte à une classe de mots générique possédant la même fonction grammaticale. Nous ne cherchons pas à remplacer les définitions déjà présentes dans le lexique car elles permettent de gérer avec précision le vocabulaire courant. Parmi les trois méthodes d'extension du vocabulaire de **LGP** évaluées par Pyysalo *et al.* [PSAN06], nous constatons que le couplage d'un étiqueteur spécialisé de *POS* est la solution est la plus performante ce qui nous conforte dans le choix de notre stratégie. Le complètement du

lexique par des ressources terminologiques externes (par exemple, le contenu du **Metathesaurus UMLS**) ou le perfectionnement des techniques de détection de caractéristiques morphologiques de surface n'ont révélé qu'un gain superficiel dans les performances de **LGP** d'après cette étude.

Une autre limite de **LGP** dans l'analyse de phrases du domaine de la biologie est la présence de groupes nominaux complexes [BRR+05]. Une étude de Pyysalo *et al.* [PGP+04] a montré que 28% des erreurs produites lors de l'analyse par **LGP** d'un corpus de biologie moléculaire étaient dues à la présence d'ENs complexes. **LGP** ne permet de relier qu'un seul nom (la *tête*) d'un groupe nominal via une conjonction de coordination, un verbe, etc au reste de la phrase. Les autres noms du groupe nominal sont alors attachés à celui-ci en tant que compléments du nom. Or, comme précédemment vu (voir la section 3.2.1.2), les ENs en biologie ne sont pas uniquement des groupes nominaux simples mais peuvent aussi contenir des conjonctions de subordination et des verbes par exemple. Le nombre de *liens* à tester devient artificiellement très grand et **LGP** présente alors de grandes difficultés à retrouver la *tête* réelle de l'EN. Nous proposons de répondre à cette contrainte lors du pré-traitement des phrases en associant à une EN identifiée un mot unique. Un *POS* spécial est alors attaché à ce mot. Ce *POS* correspond dans le lexique **LGP** à une entrée propre aux conditions d'établissement de *liens* avec une EN en biologie. La définition de cette nouvelle entrée du lexique est très proche de celle proposée pour les noms communs avec quelques caractéristiques additionnelles propres aux noms propres.

4.2 Simplification de la syntaxe et construction des structures prédicat-arguments

Comme nous l'avons précédemment vu dans la section consacrée à la représentation de la connaissance pour l'EI, nous cherchons à cette étape à structurer et à organiser la sémantique à partir du texte. Les faits exposés dans les textes et liés à un évènement d'intérêt sont déduits des relations syntaxiques que les composants de la phrase entretiennent entre eux. Néanmoins, la difficulté majeure rencontrée lors de cette structuration de l'information correspond à la diversité même de la syntaxe. En effet, un même évènement peut être décrit sous de nombreuses formes, très différentes les unes des autres. Les sources linguistiques de cette diversité sont multiples. Parmi les plus importantes, nous pouvons citer le rôle de la forme passive et de la forme active, la nominalisation des structures verbales, les tournures impersonnelles et les structures dites en *montée* (par

exemple, "IL2 receptor is expected to bind IL2" où "IL2 receptor" est le sujet sémantique de "bind").

4.2.1 Les structures d'argument

Les *prédicats argument-structures* sont une forme de représentation de l'information linguistique à mi-chemin entre la syntaxe et la sémantique [Nob88]. Ils permettent de formuler un très grand nombre de phénomènes syntaxiques et de les structurer.

Les *prédicats argument-structures* sont utiles afin d'exprimer de manière unique les différentes séquences de mots exprimant la même information. En réduisant la complexité de la syntaxe de la phrase, le temps et les efforts nécessaires au développement de règles d'extraction spécifiques au domaine en sont réduits. Par exemple, en logique des prédicats, les deux phrases suivantes : "IL2 binds IL2 Receptor" et "IL2 Receptor is bound by IL2" sont réprésentées par l'expression unique :

```
[bind](IL2, IL2 Receptor)
```

Nous ne décrirons pas la logique des propositions ni la logique de prédicats dans ce document, néanmoins nous proposons quelques définitions utiles à la suite de l'exposé. Nous pouvons ici utiliser la phrase suivante : "Seminal plasma activates cytokine genes in cervical epithelial cells, and the cytokines expressed in cervical epithelium are IL-8, IL-6, GM-CSF and MCP-1" afin d'illustrer notre propos. Cette phrase contient 4 *propositions* logiques (que nous appelons ici *définitions* et qui représentent autant d"'évènements" atomiques) : "Seminal plasma activates cytokine genes in cervical epithelial cells", "the cytokines expressed in cervical epithelium are IL-8", "the cytokines expressed in cervical epithelium are IL-6", "the cytokines expressed in cervical epithelium are GM-CSF" et "the cytokines expressed in cervical epithelium are MCP-1". Ces *définitions* correspondent aux différentes assertions (déclarations logiques) formulées dans la phrase. Chacune de ces *définitions* peut être représentées par une *structure prédicat-arguments* particulière. Les *arguments* de cette structure sont les "participants" nécessaires, mais pas toujours exprimés, pour que la situation formulée par le *prédicat* soit "complète" (pleinement réalisée). Des structures candidates de *prédicat-arguments* correspondantes aux *définitions* précédentes sont les suivantes :

```
[activate](seminal plasma, cytokine genes, cervical epithelial cells)
[expressed](IL-8, cervical epithelium)
[expressed](IL-6, cervical epithelium)
[expressed](GM-CSF, cervical epithelium)
[expressed](MCP-1, cervical epithelium)
```

Les *prédicats* sont présentés entre crochets alors que les *arguments* dont ils dépendent sont signalés entre parenthèses. D'autre part, un *prédicat* peut n'absorber qu'une information parcellaire détenue par la *définition*. Par exemple, dans la *définition* "the cytokines expressed in cervical epithelium are GM-CSF", aucun prédicat n'a été proposé au dessus afin de signifier une relation éventuelle entre "cytokines" et "GM-CSF". Dans la section 4.3.2.3 nous présentons des solutions afin de déduire des informations non formalisées grâce aux relations qui animent les *structures prédicat-arguments* répertoriées. Dans les exemples précédents, le *prédicat* 'activate' prend pour *arguments* une molécule activatrice, une molécule activée et un lieu, le *prédicat* 'expressed' lui en revanche nécessite une molécule exprimée et un lieu. L'ensemble des *arguments* sélectionnés dépendent entièrement du *prédicat* correspondant. Néanmoins, pour un même *prédicat*, certaines contraintes sur les *arguments* doivent être observées :

- *La contrainte de cardinalité.* Le nombre d'*arguments* est fixe pour une instance de *prédicat*. Dans notre exemple précédent, le *prédicat* 'activate' est toujours accompagné de trois *arguments*.
- *La contrainte d'ordre.* Par exemple, [expressed](GM-CSF, cervical epithelium) est différent de [expressed](cervical epithelium, GM-CSF). Une représentation syntaxique possible de la deuxième structure est "cervical epithelium is expressed in GM-CSF".
- *La contrainte de type.* Chaque argument est liée à une catégorie sémantique fixe. Par exemple, le premier *argument* du *prédicat* 'activate' est toujours une molécule et ne peut être un lieu ou un être humain.
- *La contrainte de formes syntaxiques.* Un *argument* ne peut appartenir qu'à un ensemble restreint de formes syntaxiques. Dans la structure exemple correspondante au *prédicat* 'expressed', le premier argument est forcément un syntagme nominal ou une EN alors que le deuxième *argument* est un syntagme prépositionnel introduit par un nombre limité de prépositions ou de conjonctions de subordination.

De part cette structuration, les *structures prédicat-arguments* sont un moyen de représenter de façon formelle les propriétés sémantiques de la langue naturelle.

4.2.2 Simplification de la syntaxe

Les listes des *liens* ainsi que les *arbres des syntagmes* proposés par **LGP** représentent une information extrêmement précise et fine sur la syntaxe de la phrase. Ceci est notamment dû à un premier apport de sémantique lors de l'analyse de la syntaxe. Les contraintes de création de *liens* imposées sur les entrées du lexique **LGP** prennent en compte certaines formes de lexicalité basées sur l'utilisation courante de la langue anglaise et donc indépendantes de tous domaines. Par exemple, le verbe 'to sell' est défini comme appartenant sémantiquement à trois catégories distinctes, régis par des règles syntaxiques strictes :

1. 'to sell' peut prendre le sens premier, c'est à dire d'échanger contre rétribution un bien à quelqu'un et peut être formalisé par la *structure prédicat-arguments* suivante :
   `[sell](qui vend, à qui, la marchandise, contre quelle somme)`
2. 'to sell' peut aussi être compris comme l'abandon d'obligations en échange d'avantages comme dans la phrase "The robber sold his accomplices out to the police".
   `[sell](qui vend, à qui, dans le cadre de quelle obligation, contre quel avantage)`
3. 'to sell' peut aussi finalement convoyer le sens de persuasion comme dans l'exemple "John sold his project to the commitee"
   `[sell](qui vend, à qui, quel projet)`

Il est à noter que dans les exemples de *structures prédicat-arguments* présentés ci-dessus, certains arguments sont facultatifs.

Nous nous proposons dans cette section d'utiliser cette première injection de sémantique au travers de *structures prédicat-arguments* génériques afin de simplifier la syntaxe. Notre but ici est uniquement de réorganiser la syntaxe et de gommer ses différentes représentations. Aucune considération lexicale dépendante du domaine de la régulation de gènes n'est encore envisagée. Nous avons sélectionné deux *structures prédicat-arguments* prédéfinies afin de représenter l'ensemble de la diversité syntaxique d'une phrase. En réduisant la complexité de la syntaxe de la phrase, le temps passé à l'élaboration les règles d'extraction spécifiques au domaine est réduit. De plus, moins d'exemples issus des données d'entrainement seront nécessaires.

Nous avons décidé de résumer et de décrire chaque évènement présenté dans une phrase autour de l'information convoyée par le verbe. La *structure prédicat-arguments* retenue (voir la figure 4.2) est composée de plusieurs éléments, parfois optionnels : une ACTION à laquelle peut être liée un ACTEUR (initiateur de l'action), un ACTÉ (objet de l'action)

et différentes CONDITIONs (celles ci représentent l'environnement de l'ACTION sous la forme de données complémentaires). De la même manière, certaines CONDITIONs sont spécifiques à un ACTEUR ou un ACTÉ et non à l'ACTION directement. Un élément CONDITION peut compléter un autre élément avec une information supplémentaire ou au contraire déclencher un évènement conditionnel. Le *prédicat* générique est ici l'élément que nous appelons ACTION alors que ses arguments constituent tous les autres éléments qui lui sont attachés. Le lien qui unit une CONDITION à son ACTION est explicitement spécifié dans la structure car il précise la connexion logique entre les deux éléments (par exemple, dans les deux expressions suivantes constituées de deux structures : "IL-2 is activated after TCR is stimulated" et "TCR is stimulated before IL-2 is activated", le lien souligné permet d'une part de qualifier et d'autre part d'orienter la relation qui anime les deux structures).

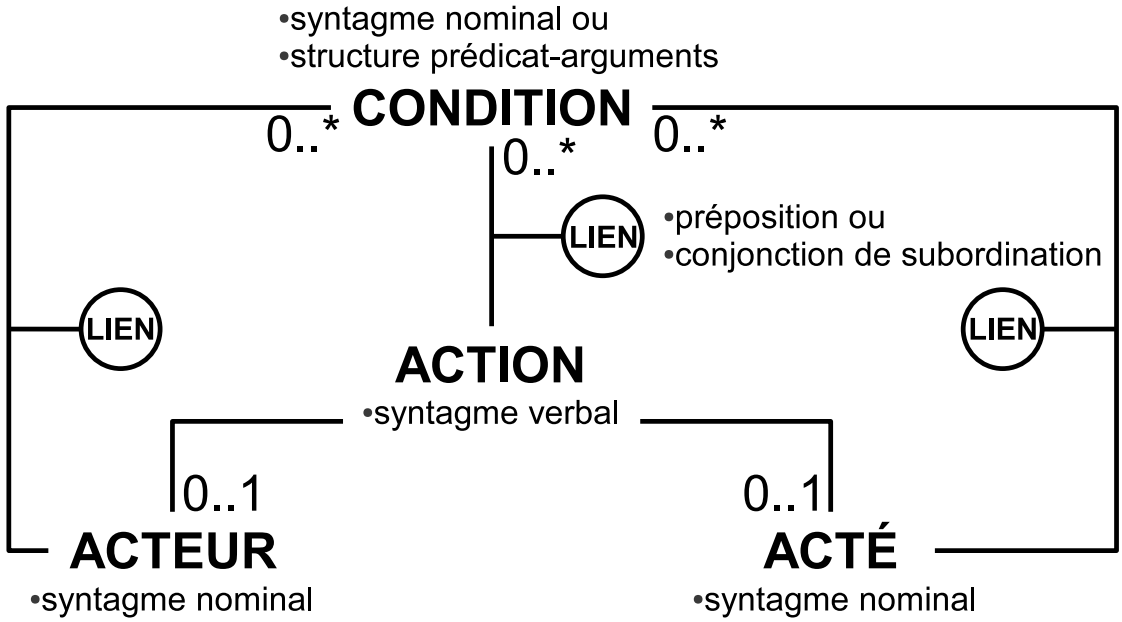


Fig. 4.2 – Structure prédicat-arguments générique

La présence de phrases (et non juste de propositions) entièrement nominalisées, c'est à dire lorsqu'aucun verbe n'est présent, est une situation particulière que nous ne prenons pas en charge et qui n'est donc pas formalisée dans une *structure prédicat-arguments*. Nous avons observé que ce type de phrases est limité aux titres des sections d'un document scientifique et est un phénomène relativement peu fréquent. Il est aussi important de noter que **LGP** est incapable de déterminer la syntaxe de telles phrases.

La situation est différente lorsque l'élément *action* contient un verbe 'transparent', c'est à dire qu'il ne correspond pas au *prédicat*, la véritable *action* étant contenue dans un autre élément sous une forme nominalisée. Par exemple, dans la phrase "we *detected* the activation of IL-2 by the TCR", le verbe *action* 'transparent' est ici "detected" alors que la véritable *action* est "activation". Ce phénomène n'est pas le propos de cette partie sera traitée dans une section ultérieure.

Les paragraphes suivants détaillent les différentes méthodes nécessaires à la création de *structures prédicat-arguments* génériques à partir des listes des *liens* et des *arbres de syntagmes* **LGP**.

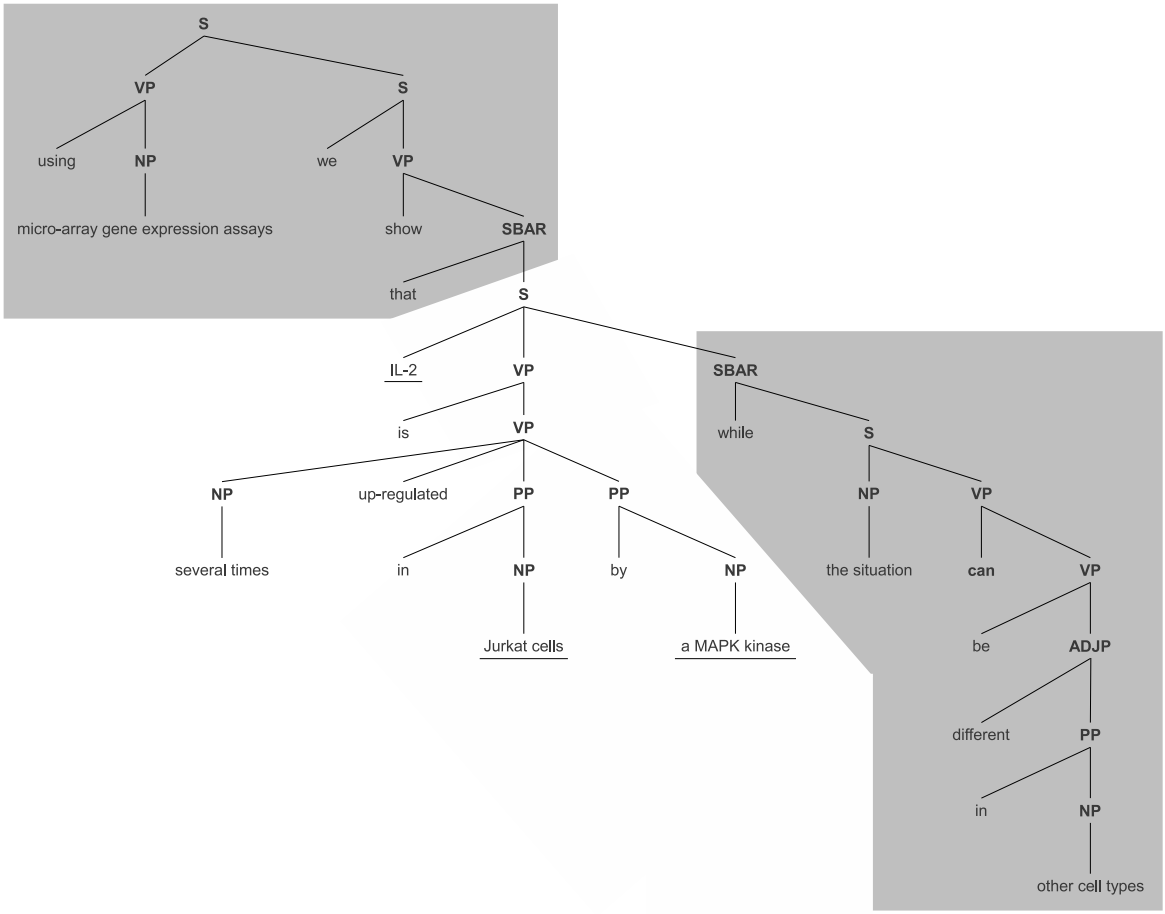
4.2.2.1 Focalisation du discours

Dans le but de limiter le nombre d'opérations et de traitements à effectuer, nous ne nous intéressons uniquement qu'aux portions de texte pouvant éventuellement décrire une relation entre deux ENs. A partir du corpus décrit dans la section 4.4.1, nous avons estimé que le temps de calcul global de la recherche de relations entre ENs pouvait diminuer de près de 5 à 10% en procédant à l'analyse de ces seuls segments. Parallèlement, nous n'avons observé aucune fluctuation significative des performances brutes du système lorsque nous procédons à la seule analyse de ces portions de texte et non des phrases entières.

A partir de l'*arbre des syntagmes* **LGP** nous ne conservons que le texte correspondant au chemin permettant de joindre l'ensemble des ENs de la *définition*. Nous élaguons les branches de l'arbre qui ne contiennent pas d'ENs sous certaines conditions. Un exemple d'élagage est proposé dans la figure 4.3 où est figuré l'*arbre des syntagmes* **LGP** correspondant à la phrase "Using micro-array gene expression assays, we show that IL-2 is several times up-regulated in Jurkat cells by a MAPK kinase while the situation can be different in other cell types". Les syntagmes étiquetés `NP` correspondent à des groupes nominaux, `VP` à des groupes verbaux, `ADJP` à des groupes adjectivaux, `PP` à des locutions prépositionnelles, `S` et `SBAR` à différentes formes de propositions (subordonnées, relatives, indépendantes, etc). Les nœuds et feuilles présentés en grisé sont excisés de l'arbre. L'arbre résultant de l'opération d'élagage correspond à la portion de texte : "IL-2 is several times up-regulated in Jurkat cells by a MAPK kinase".

Nous avons choisi de ne pas nous limiter au chemin le plus court de part la structure même des arbres **LGP**. A un même nœud peuvent être rattachés différents syntagmes

Fig. 4.3 – Exemple d'élagage d'un *arbre des syntagmes* **LGP**



Les ENs d'intérêt sont soulignées. Les zones gisées correspondent aux portions du texte supprimé. Dans un souci de simplification, les noms des syntagmes qui ne contiennent qu'un seul mot ne sont pas figurés.

essentiels à la structuration de celui-ci. Les ramifications issues de ces nœuds fils peuvent néanmoins ne pas contenir d'ENs. Par exemple, les syntagmes verbaux et leurs syntagmes nominaux sujets descendent directement du même nœud et n'entretiennent donc pas de relation père-fils. Nous élaguons les portions de l'arbre des feuilles jusqu'aux nœuds décrivant la clauses propositionnelle la plus étendue et en l'absence d'ENs. De plus, la racine de l'arbre est déplacée au premier nœud contenant une EN. Dans l'exemple proposé plus haut, le syntagme verbal correspondant au texte "IL-2 is several times up-regulated in Jurkat cells by a MAP kinase" correspond au nouvel *arbre des syntagmes* une fois réduit. La phrase est syntaxiquement correcte et conserve les informations relatives aux relations animant les ENs d'intérêt.

4.2.2.2 Construction des structures prédicat-arguments génériques

Ce paragraphe présente les différentes procédures utilisées afin de ré-associer les différents syntagmes de l'*arbre des syntagme* produit par **LGP** aux éléments des *structures prédicat-arguments* génériques.

Les différents syntagmes sont réorganisés afin de correspondre aux contraintes imposées par les *structures prédicat-arguments*. Le processus d'association consiste, de manière figurée, à couper l'*arbre des syntagmes* **LGP** au nœud correspondant au syntagme d'intérêt et à déplacer ce dernier dans un élément particulier d'une structure. Il est à noter que le syntagme excisé de l'arbre contient toutes ses ramifications et donc les syntagmes qu'il encapsule. A son tour, le syntagme enregistré dans un élément peut être re-découpé et un des produits de cette césure ré-associé à un autre élément.

L'ensemble de ces opérations de construction des *structures prédicat-arguments* est régi par des règles basées sur les *liens* **LGP** que les syntagmes entretiennent entre eux.

La liste des *liens* **LGP** utilisée dans le cadre de l'élaboration de ces règles ainsi qu'une courte description sont proposées dans le tableau 4.1. Les détails sont accessibles à partir de la documentation officielle de **LGP**[3].

1. *Création des squelettes de structures prédicat-arguments et isolement des syntagmes verbaux.*

   Nous cherchons à affiner les *définitions* proposées par **LGP** qui, nous le rappelons, sont générées lorsqu'une énumération est présente dans la phrase. Chaque *définition* **LGP** est décomposée en sous-*définitions* atomiques, basées sur la présence d'un verbe central et compatible avec notre *structure prédicat-arguments*.

   Dans le reste du paragraphe, chaque ensemble de structures issu d'une *définition* **LGP** originelle sera évalué en parallèle.

   Les syntagmes verbaux (`VP`) sont isolés de l'arbre et permettent de définir autant de squelettes de *structures prédicat-arguments* (chaque élément ACTION d'une structure étant composé d'un syntagme `VP`).

   Les relations qui animent les structures entre elles (via l'utilisation d'un élément CONDITION prévu par la structure) seront évaluées plus tard, lorsque chaque structure sera complète et valide.

[3] http ://www.link.cs.cmu.edu/link/dict/summarize-links.html

Tab. 4.1 – Liens LGP d'intérêt pour la détermination des éléments d'une structure

Lien	Description succincte	Exemple
S	Connecte les noms et gérondifs sujets aux verbes 'finis'	Playing the piano is fun The dog chased the cat
SI	La forme interrogative de S	Is Jane coming
SF	Connecte les sujets apparents qui ne représentent rien ("it" et "there") aux verbes impersonnels	It seems that Jane should go
SFI	La forme interrogative de SF	Is there going to be a problem
SX	Connecte "I" à "was" et "am"	I am gone
SXI	La forme interrogative de SX	Am I gone
Mg	Connecte les noms aux participes présents	The dog chasing the man died
MX	Connecte les noms aux compléments du nom placés entre des virgules	The dog, with a big nose, barked loudly The dog, chasing the cat, barked loudly
Mr	Connecte les noms aux propositions relatives possessives	the dog whose owner died was black
Mv	Connecte les noms aux participes passés	The dog chased by the man died
O	Connecte les verbes transitifs aux objets directs et indirects	The dog chased the cat
OD	Connecte des verbes tels que "rise" et "fall" à des expressions de distance compléments	It fell five feet
OT	Connecte des verbes tels que "last" à des expressions de temps compléments	It lasted five years
OF	Connecte certains verbes et adjectifs au mot "of"	I thought of something I am proud of you They accused him of the crime
OX	Connecte certains objets aux verbes fonctionnant avec "to" et dont le sujet est apparent	I expected it to be easy
B	Connecte certains verbes à leurs objets lorsque ceux ci sont situés à leur gauche et non à leur droite	The dog I had chased was black
TOo	Connecte un verbe à un complément à l'infinitif	I advised him to go
TOt	Connecte certains adjectifs avec des constructions infinitives transitives	John is easy to hit
MVi	Connecte des constructions à l'infinitif aux verbes et adjectifs lorsqu'ils prennent le sens de "in order to"	He went to the store to get some bread
MVx	Connecte les verbes aux propositions basées sur un participe à la forme passive ou progressive et entourées par de virgules	John left, carrying a dog
MVp	Connecte les verbes à certains participes ou prépositions	She prepared for the meeting She cried when asked about it
MVs	Connecte les verbes aux conjonctions telles que "while", "because" et "after"	John left the party now that Jim escaped
I	Connecte les verbes aux infinitifs	I made him go
CO	Connecte les propositions d"'ouverture" aux sujets des propositions principales	On tuesday, they went out Although they were tired, they went out Leaving the kids at home, they went out
R	Connecte les noms aux propositions relatives	The dog who chased me was black

Dans les exemples proposés, le lien est établi entre les deux mots soulignés.

2. *Identification des éléments ACTEURs d'une structure.*

   Un *élément* ACTEUR est associé à chaque ACTION.
   - L'ACTEUR est le syntagme qui contient un mot proposant un *lien* **S**, **SF**, **SFI**, **SI**, **SX** et **SXI**, puis à défaut **Mg** et enfin **CO** avec un mot du syntagme de ACTION.
   - Dans le cas où la *tête* de l'ACTEUR est un mot tel que 'who', 'that', 'which', etc et entretient un lien **MX** avec un autre syntagme, l'ACTEUR est remplacé par ce dernier.
   - Les éléments ACTEURs 'whose' et 'which', qui introduisent des propositions relatives possessives, sont remplacés par les *têtes* des syntagmes principaux qu'ils décrivent grâce aux liens **Mr**. Seul le pronom possessif est substitué et non le syntagme entier. Pour illustration, la proposition soulignée dans l'exemple "IL2 receptors which function has been assessed by immuno-fluorescence are detected on the membrane cell" a pour ACTEUR "IL2 receptors function" après transformation et non "IL2 receptors" uniquement.
   - Il est à noter que certaines constructions particulières impliquent une relation sujet-verbe avec un *lien* **B**. Ces situations sont similaires à l'emploi d'un lien **MX** mais sans la présence de virgules (par exemple, dans la phrase "The protein which has been expressed is detected", les mots soulignés sont connectés grâce à un *lien* **B**).

3. *Identification des éléments ACTÉs d'une structure.*

   Un *élément* ACTÉ est associé à chaque ACTION.
   - L'ACTÉ est le syntagme qui contient un mot proposant un *lien* **Mv**, **O**, **OD**, **OT** et **OX**, puis à défaut **OF** avec un mot du syntagme de ACTION. La présence d'un *lien* **Mv** implique que la première CONDITION dépendante de l'ACTÉ introduite par une préposition 'by' devient automatiquement l'ACTEUR de la *définition*. Dans certaines formes passives, d'autres prépositions telles que 'through', 'via' ou 'upon' peuvent être indifféremment substituées à 'by'. Néanmoins ces cas particuliers sont spécifiques au verbe de l'ACTION et sont alors gérés ultérieurement, lors de la conceptualisation (voir la section 4.3.2.2).
   - Dans le cas où la *tête* de l'ACTÉ est un mot tel que 'who', 'that', 'which', etc et entretient un lien **B** avec un autre syntagme, l'ACTÉ est remplacé par ce dernier.
   - Les ACTÉs 'whose' et 'which', qui introduisent des propositions relatives possessives, sont remplacés par les *têtes* des syntagmes principaux qu'ils décrivent grâce aux liens **Mr**. De la même manière que pour les ACTEURs, seul le pronom possessif est substitué et non le syntagme entier.

4. *Prise en compte des éléments ACTÉs ou ACTEURs vides.* A ce stade, certains éléments ACTÉs ou ACTEURs n'ont pas été renseignés car leur association avec une ACTION n'est pas syntaxiquement explicite. Leur résolution impliquent la prise en considération de constructions verbales imbriquées où un même syntagme est lié simultanément à un élément ACTÉ ou ACTEUR de deux ACTIONs. Ce syntagme peut alors remplir la même fonction ACTÉ ou ACTEUR pour les deux structures ou au contraire tenir un rôle opposé. Les règles suivantes permettent de spécifier un élément ACTÉ ou ACTEUR non encore identifié à partir d'autres éléments préalablement découverts. Un élément déjà spécifié grâce aux règles présentées ci-dessus n'est pas remplacé.
   - Dans le cas où deux syntagmes verbaux et donc deux structures sont connectées par un *lien* **TOo**, l'ACTÉ de la structure correspondant au syntagme de gauche est aussi l'ACTEUR de la structure du syntagme à droite.
   - Si deux structures sont unies par un *lien* **TOt**, **MVi**, **MVp**, **MVx**, **MVs** et **I**, les ACTEURs des deux structures sont identiques.
   - Avec les *liens* de type **Mv**, **Mg** et **MX**, la situation est plus contrastée. En règle générale l'ACTEUR est partagé entre les deux structures réunies grâce à un de ces *liens*. Néanmoins, il existe de nombreux contre-exemples (pour illustration, "We angered her by chasing the dog" et "She criticized us for chasing the dog" partagent le même parallélisme de construction des *liens*, pourtant le sujet de "chasing" est soit "we" soit "us" selon la phrase). Nous ne prenons en compte que la règle générale et admettons que certaines erreurs d'assignation peuvent subvenir sans pour autant les quantifier.
5. *Identification des éléments CONDITIONs d'une structure.*
   - *Les groupes nominaux CONDITIONs.* Un ou plusieurs *éléments* CONDITIONs sont associés à une ACTION, un ACTEUR ou un ACTÉ selon les syntagmes unis par les *liens*. Certaines CONDITIONs peuvent être connectées en *cascade*, c'est à dire qu'un syntagme définissant une CONDITION est relié à une autre CONDITION et non à un ACTEUR, ACTÉ ou ACTION directement. Dans ce cas nous re-connectons chaque CONDITION de cette séquence, individuellement, à l'élément ACTEUR, ACTÉ ou ACTION sur laquelle la séquence est branchée. Nous conservons la notion d'ordre dans la séquence originale lors de la ré-association. La notion de CONDITION est très large, nous considérons toutes les locutions prépositionnelles (PP) à base de groupes nominaux comme des instances valides d'éléments CONDITIONs.
   - *Les structures CONDITIONs et la détermination des relations entre structures.*

Comme souligné précédemment, les *structures prédicat-arguments* codent les différentes assertions logiques du discours. Cette information est basée sur l'utilisation d'un verbe central qui supporte une ACTION. La mise en relation des structures peut être utile lorsque l'information convoyée par une structure unique est soit parcellaire soit soumise à certaines réserves. Par exemple, dans la phrase "The expression of the CD95 ligand is elevated when T cells are activated", l'information sémantique portée par les deux structures correspondant d'une part au syntagme "The expression of the CD95 ligand is elevated" et d'autre part "T cells are activated" n'est exploitable que dans le cadre de leur confrontation lors d'une situation particulière. Alors que les liens potentiellement établis entre différentes phrases ne sont pas explicités par les auteurs du documents, les liens présents au sein d'une même phrase peuvent être précisés grâce à la syntaxe. Il est aussi intéressant de noter que ces liens sont orientés. Par exemple, la sémantique derrière ces deux phrases : "The expression of the CD95 ligand is elevated when T cells are activated" et "T cells are activated when the expression of the CD95 ligand is elevated" est différente.

Nous créons un lien de dépendance entre deux structures lorsqu'un *lien* **LGP** permet d'unir un des syntagmes utilisés par chaque structure. Néanmoins, dans le cadre d'une CONDITION, nous ne nous intéressons qu'aux relations se rapportant à une subordination circonstancielle ou relative. Les liens d'intérêt retenus sont **R**, **CO**, **MX** et **MV**. Les autres types de relations ne sont pas détectées. La structure spécifiée à droite du *lien* est alors incluse dans un élément CONDITION dépendant de l'ACTION de la structure à gauche. Il existe quelques exceptions à cette règle, notamment dans le cas des *liens* **CO** où l'ordre est inversé. Les conjonctions de subordination (par exemple, "when", "where", "while", "although") et certains type de pronoms relatifs sont très importants car ils définissent la nature de la relation et le lien logique qui unit les deux structures et sont donc explicitement conservés au sein de l'élément CONDITION.

6. *Regroupement en noms composés.* D'autre part, nous procédons à la fusion de certains éléments lorsqu'ils peuvent donner naissance à un nom composé. Nous choisissons de ne garder qu'une seule représentation syntaxique au sein des structures pour ce type de groupes nominaux particuliers. Les CONDITIONs dépendantes d'un ACTEUR ou d'un ACTÉ sont parcourues dans le sens inverse de leur stockage. Un élément (*n*) introduit par "of" est fusionné avec l'élément suivant (*n-1* dans l'ordre séquentiel d'enregistrement des éléments CONDITIONs) afin de former un nom composé. La méthode de fusion est identique à celle proposée dans la

première étape de normalisation exposée dans le paragraphe 3.1.2.3, à la différence que la préposition qui introduit l'élément *n-1* est conservée. Le nouveau élément issu de cette combinaison sert alors de base à l'élaboration d'un autre nom composé avec l'élément suivant au besoin. Si le dernier élément CONDITION de la liste (le premier dans l'ordre de stockage) est introduit par "of", il est intégré à l'élément ACTEUR ou ACTÉ dont il dépend sous la forme d'un nom composé.

4.2.2.3 Quantification et modulation de l'information contenue dans les structures

En l'état actuel de notre système d'identification des relations biologiques d'intérêt, nous ne gérons pas la quantification et certaines formes de modulation des relations extraites. Certaines subtilités telles que les différences sémantiques entre les expressions "half of the cytokines" et "all cytokines" ou "is mildly expressed" et "is greatly expressed" ne sont pas prise en considération. Nous nous positionnons dans une approche binaire d'extraction de l'information : les données présentes et exploitables sont considérées comme absolues et valeurs de vérité. Néanmoins, nous prenons en considération certaines formes d'incertitude et le phénomène de négation.

1. Le doute quant à la véracité d'une proposition émise est courant dans les documents expérimentaux en biologie. Il est d'usage pour un scientifique de proposer alternativement des résultats scientifiques, considérés comme des faits objectifs, et des hypothèses de travail et des interprétations de résultats non vérifiées dans ses articles. De la même manière, nous considérons deux degrés uniquement : soit la proposition correspond à une vérité établie, soit c'est une hypothèse. Nous avons à notre disposition une liste d'adverbes (par exemple, "arguably", "unlikely", "hardly"), d'adjectifs (par exemple, "independent", "failed", "defective", "false", "insufficient", "counterfactual"), des modaux (par exemple, "may", "might") et de verbes qui peuvent être indifféremment retrouvés sous des formes adjectivales (par exemple, "hypothesize", "mislead", "unverify", "suspect"). Nous cherchons dans les éléments des structures leur présence. Si un de ces mots a été détecté, la structure entière et les structures qu'elle coordonne sont marquées comme étant incertains. ce marquage est utile lors de l'agrégation finale des données d'un document et dans la résolution des données contradictoires à la section 4.3.3.
2. Les opérateurs linguistiques de négation inversent le sens d'une proposition et doivent être pris en compte. Bien que les négations puissent être informatives (par exemple,

la protéine X n'active pas le gène Y), nous ne conservons que les propositions affirmatives pour analyse. Par exemple, l'information que "IL3 does not activate IL4" n'est pas utilisable dans notre modèle de la régulation de gènes. Les mots symbolisant la négation absolue (tels que "no", "not" et "never") sont recherchés dans les éléments de chaque structure. Lorsqu'un nombre impair de négations est retrouvé dans une structure, cette dernière ainsi que toutes les structures qu'elle coordonne sont supprimées et ne seront pas donc pas analysées plus avant. Nous partons du principe que deux négations s'annulent au sein d'une même structure. Ainsi, la structure correspondant à la phrase exemple "**No** IL2 does **not** bind to the IL2 receptor" est équivalente à la phrase "IL2 binds to the IL2 receptor". Néanmoins la présence de deux négations ou plus dans une proposition est rare dans les articles biomédicaux. Il est à noter que certains idiomes tels que "not to forget" ou "last but not least" sont préalablement identifiés et ne sont pas considérés comme des marques de négation. Une des limites de cette approche simpliste et globale de la gestion des négations est la perte des informations contenues dans les éléments CONDITIONs lorsqu'une négation est présente. Par exemple, dans la phrase "IL2-stimulated NK cells up-regulates TNF-α when **no** anti-LFA-1 integrin mAb is injected", la structure correspondant à la proposition soulignée est éliminée alors qu'elle est essentielle à la définition de la proposition principale. Les cellules NK stimulées par l'IL2 ne régulent positivement le TNF-α que lorsqu'aucun anticorps monoclonal dirigé contre l'intégrine LFA-1 n'est présent. Dans le contexte de la biologie expérimentale, cela signifie que les cellules NK stimulées par l'IL2 ne régulent positivement le TNF-α qu'avec l'aide de l'intégrine LFA-1. Afin de s'affranchir de cette limitation, la résolution du rôle des négation pourrait être adressée plus tard, à l'étape de création de règles d'extraction des relations entre ENs, en contrepartie d'une plus grande complexité du jeu de règles développées. Ce point reste à explorer.

4.2.2.4 Simplification des éléments des structures

Les éléments des *structures prédicat-arguments* contiennent des expressions textuelles qui sont à la base du système de règles et doivent encore être formatés.

- Le texte contenu dans chaque élément d'une structure est lemmatisé. Nous utilisons une version expurgée de l'algorithme de **Porter** (à la différence de celle présentée dans la dernière partie du paragraphe 3.1.2.3) afin de ne s'affranchir que des marques du pluriel et des temps. En effet, certains suffixes seront activement utilisés lors la

phase de conceptualisation des structures alors que d'autres sont soit inutiles ou facilement retrouvables grâce au *POS* qui leur est associé.

- D'autre part, certains termes sont supprimés car ils ne sont soit pas utilisables par le système ou ont été déjà pris en considération dans la structure de façon non textuelle (comme par exemple certains modulateurs et quantificateurs liés à l'incertitude). Ces sont les déterminants, les expressions adverbiales, les modaux et les verbes auxiliaires "to have" ou "to be" utilisés en conjonctions avec les formes progressives et le *present perfect*. Ces ensembles de termes sont détectés grâce à un jeu de règles utilisant à la fois les *liens* **LGP** et les *POS*.

Dans le cas où une structure se retrouvait sans ACTION (c'est notamment le cas des structures dont l'ACTION est l'auxiliaire "to have" ou "to be" des temps composés), la structure entière serait supprimée.

4.2.2.5 Exemples de structures prédicat-arguments

Les figures 4.4 et 4.5 présentent deux illustrations de transformation d'un arbre de syntagmes **LGP** en *structures prédicat-arguments* grâce à l'ensemble des règles proposées ci-dessus. Dans un souci de simplification, les noms de certains syntagmes de l'arbre qui ne contiennent qu'un seul mot ne sont pas figurés.

A partir de ces exemples nous montrons que certaines représentations syntaxiques différentes ont pu être absorbées grâce aux structures *prédicat-arguments*.

1. L'AUTEUR et l'ACTÉ des ACTIONs sont correctement détectés.
   (a) Les formes passives et actives sont structurées. Par exemple, dans la figure 4.4, "MAP kinase" est bien l'ACTEUR de l'ACTION "up-regulate" et non "IL2".
   (b) Certaines formes d'anaphores simples sont résolues. Par exemple, dans la figure 4.5, "which activation" a été interprété en tant que "JAK1 activation" grâce à la règle spécifique au cas des possessifs.
2. les formes syntaxiques en *montée* sont interprétées. Dans la figure 4.5, le syntagme "We show that" a donné naissance une structure isolée car il n'établit pas de liens d'intérêt avec les autres syntagmes et sera ignorée par les règles spécifiques au domaine d'extraction des relations.
3. les problèmes de morphologie et de temps sont pris en considération. Dans la figure 4.4, "was up-regulated" a été scindé en deux ACTIONs distinctes, la structure basée

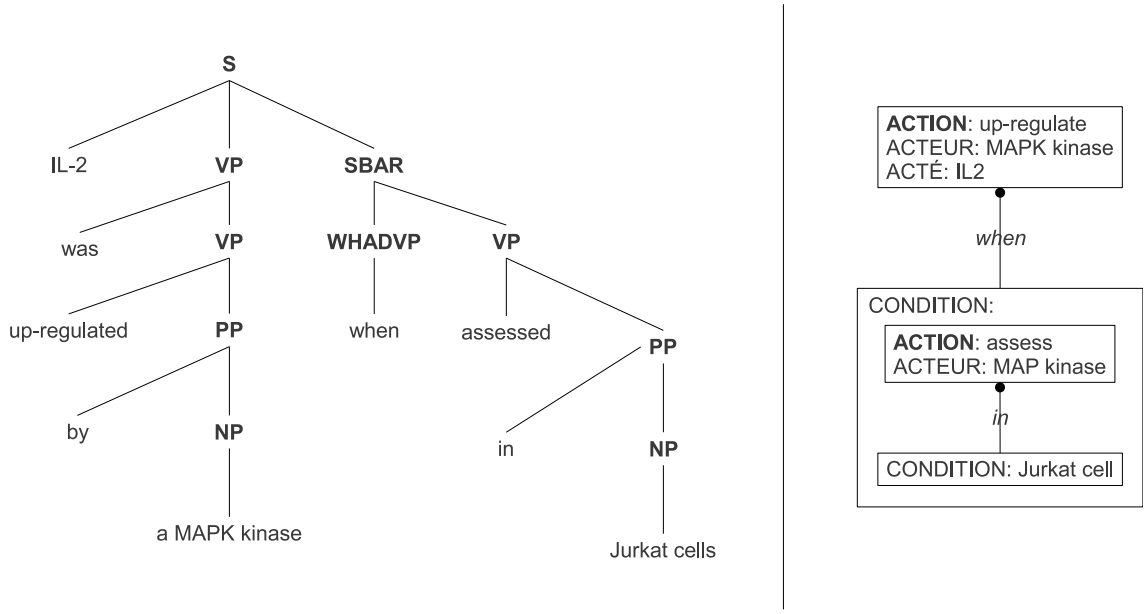


Fig. 4.4 – Exemple d'un *arbre des syntagmes* **LGP** (à gauche) et de sa représentation en *structures prédicat-arguments* (à droite)

sur "was" a été supprimée grâce aux règles spécifiques à l'usage des auxiliaires des temps composés.

4.3 Des structures prédicat-arguments génériques aux schémas de cas spécifiques de la régulation de gènes

Afin de découvrir les relations fonctionnelles qui unissent les entités biologiques d'intérêt que sont les ENs, identifiées grâce aux méthodes exposées dans la section 3.2, nous employons une stratégie dite *ascendante* d'injection de la sémantique à partir des structures d'abstraction de la syntaxe que sont les *structures prédicat-arguments*.

La représentation du texte en *structures prédicat-arguments* génériques nous permet de réduire significativement le nombre de différentes représentations de la syntaxe et d'organiser la structure de la phrase. Néanmoins, le niveau d'abstraction de la phrase est encore insuffisant pour se résoudre à utiliser des techniques de recherches de motifs propres à chaque relation biologique d'intérêt que nous nous efforçons d'isoler des textes.

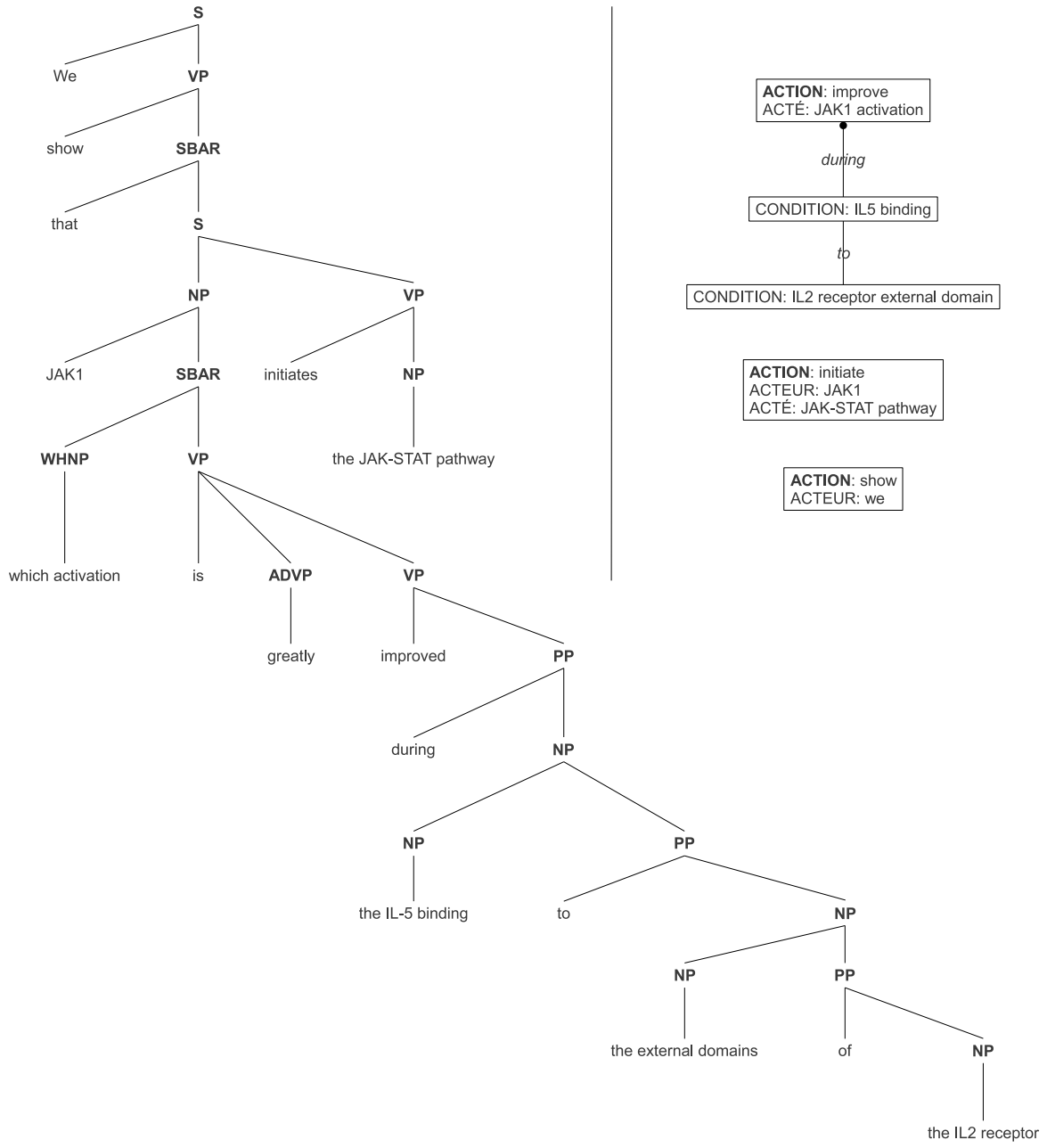


Fig. 4.5 – Exemple d'un *arbre des syntagmes* **LGP** (à gauche) et de sa représentation en *structures prédicat-arguments* (à droite)

D'une part les questions relatives au vocabulaire restent ouvertes. La terminologie du domaine de la biologie moléculaire est riche et les combinaisons nombreuses pour exprimer la même information. Par exemple, l'expression "The α and β domains of the IL2 receptor interacts with IL2 " est strictement équivalente à "The α and β regions of the IL2 receptor binds IL2".

D'autre part, certaines difficultés d'ordre purement syntaxiques demeurent. L'exemple suivant implique un phénomène linguistique commun appelé **alternance d'arguments internes**. Le verbe "to bind" en biologie moléculaire et en biochimie peut exprimer ses arguments de deux façons différentes au minimum, c'est à dire "IL2 **binds** the α and β chains *on* IL2 receptor" et "IL2 receptor **binds** IL2 *with* the α and β chains". Sémantiquement équivalentes, ces deux phrases sont néanmoins représentées par deux structures distinctes. Le deuxième exemple est propre au procédé linguistique de nominalisation des verbes et qui consiste à utiliser un verbe en tant que nom. Ce dernier devient alors la *tête* du groupe nominal. Par exemple, les expressions "We detected the **activation** of IL2 upon TCR-mediated stimulation" et "We detected that IL2 **is activated** upon TCR-mediated stimulation" sont sémantiquement équivalentes mais représentées par des structures bien distinctes. La première étant la forme nominalisée de la seconde.

Nous pensons que la résolution de ces différentes difficultés peut être réalisée grâce à l'apport de connaissances propres au domaine de l'étude, par étapes successives de conceptualisation des données proposées par les *structures prédicat-arguments* génériques.

4.3.1 Les concepts

Les concepts correspondent à des classes d'objets qui permettent d'abstraire et de généraliser plusieurs expressions d'une même notion.

Nous avons créé une hiérarchie de concepts propres au domaine de l'étude, en l'occurrence la régulation de gènes telle que définie dans la section 2.5.

Les concepts de plus bas niveau correspondent aux briques élémentaires de notre modèle de la régulation de gènes. Ils peuvent être combinés afin de donner naissance à de nouveaux concepts, de niveau d'abstraction plus élevé, et qui représentent sémantiquement le résultat de l'union des sous-concepts qui les composent. La façon dont les concepts sont agrégés et les conditions de validité de leur union sont gouvernées par un ensemble

de règles, présenté dans la section suivante.

Les concepts du niveau d'abstraction le plus haut correspondent aux entités et aux relations qui les unissent au sein de notre modèle de la régulation de gènes. La liste des concepts de niveau le plus haut retenus est proposée dans la table 4.6. Ces concepts correspondent au final aux relations d'intérêt entre ENs à extraire des textes et sont présents dans notre modèle de la régulation décrit dans la figure 2.9. Il est important de noter que nous avons ajouté certains concepts représentant soit des ellipses ou des raccourcis soit des relations non-déterminées parmi les relations compatibles avec notre modèle de la régulation. Leur présence est essentielle dans l'élucidation des instances de notre modèle de la régulation (voir la partie consacrée aux difficultés liées aux données manquantes d'un modèle dans la section 2.5). Par exemple, les concepts liés à la notion de *régulation* peuvent implicitement spécifier soit une *activation* soit une *répression*, néanmoins les auteurs de l'article ne lèvent pas le voile sur cette ambigüité dans le texte. Cette imprécision peut être soit réellement présente dans le texte (par exemple, les auteurs ne le savent pas eux même ou cette régulation cache un mécanisme mixte d'activation et de répression) soit trop complexe pour être désambigüisée par le système (par exemple, la phosphorylation ou la méthylation d'une protéine peuvent tout aussi bien être synonymes d'activation comme de désactivation sans plus d'indication).

Des exemples de concepts de plus bas niveau et de concepts intermédiaires sont présentés dans la section suivante.

Il est à noter que les concepts découverts à partir des textes sont automatiquement instanciés s'ils impliquent des ENs. Lorsqu'un concept est identifié, les ENs qui correspondent à cette portion de texte sont associées à cette occurrence particulière du concept. Les concepts sont définis de telle manière que les références aux ENs correspondantes sont structurées : les rôles des ENs joués au sein du concept sont calqués sur ceux des éléments des *structures prédicat-arguments* qui les contiennent. Ces fonctions sont spécifiquement attribuées selon le concept, un concept particulier peut ainsi nécessiter qu'une EN prenne un rôle similaire à celui d'ACTEUR, d'ACTÉ ou de CONDITION afin d'être instancié. Par exemple, un concept *activation d'une gène dans une cellule par une protéine* ne peut être associé qu'à trois ENs : une de la classe 'gène' qui prend le rôle d'ACTEUR, une de la classe 'cellule' assimilée à une CONDITION et la dernière, de la classe 'protéine', qui est semblable à un ACTÉ.

Dans la suite du document, les noms (ou labels) des concepts sont donnés sous forme

Fig. 4.6 – Concepts de niveau le plus haut et relations d'intérêt extraites entre ENs.

Dépendance du contexte cellulaire	Relation ternaire	Relation elliptique	Nature de la relation spécifiée	l'activation de la transcription d'un gène dans une cellule par une protéine
				la répression de la transcription d'un gène dans une cellule par une protéine
			Nature de la relation non spécifiée	la régulation de la transcription d'un gène dans une cellule par une protéine
		Relation directe	Nature de la relation spécifiée	l'activation de la transcription d'un gène dans une cellule par un facteur de transcription
				la répression de la transcription d'un gène dans une cellule par un facteur de transcription
				l'activation de la transcription d'un gène dans une cellule par un site de liaison aux facteurs de transcription
				la répression de la transcription d'un gène dans une cellule par un site de liaison aux facteurs de transcription
			Nature de la relation non spécifiée	la régulation de la transcription d'un gène dans une cellule par un facteur de transcription
				l'interaction entre un facteur de transcription et un site de liaison aux facteurs de transcription dans une cellule
	Relation binaire	Relation elliptique	Nature de la relation spécifiée	l'activation de la transcription d'un gène par une protéine
				la répression de la transcription d'un gène par une protéine
			Nature de la relation non spécifiée	la régulation de la transcription d'un gène par une protéine
		Relation directe	Nature de la relation spécifiée	l'activation de la transcription d'un gène par un facteur de transcription
				la répression de la transcription d'un gène par un facteur de transcription
				l'activation de la transcription d'un gène par un site de liaison aux facteurs de transcription
				la répression de la transcription d'un gène par un site de liaison aux facteurs de transcription
			Nature de la relation non spécifiée	la régulation de la transcription d'un gène par un facteur de transcription
				l'interaction entre un facteur de transcription et un site de liaison aux facteurs de transcription
Indépendance du contexte cellulaire				la présence d'un site de liaison aux facteurs de transcription dans le promoteur d'un gène
				la localisation (mesurée en paires de bases) d'un site de liaison aux facteurs de transcription dans le promoteur d'un gène
				la composition (en acides nucléiques) d'un site de liaison aux facteurs de transcription dans le promoteur d'un gène

Dans un souci de simplification, l'ensemble des relations présentes dans notre modèle (voir la section 2.5) n'est pas extrait des textes.
Nous nous efforçons de n'extraire que les relations, ainsi que certaines propriétés annexes qui leur sont associées, que nous considérons comme clefs.

mnémonique afin de faciliter d'une part leur compréhension et d'autre part leur manipulation.

4.3.2 Les règles d'agrégation des concepts

Une technique *ascendante* est utilisée afin d'agréger les concepts jusqu'à ce que les relations cibles décrites dans la figure 4.6 soient identifiées.

Les concepts ne sont pas inférés automatiquement mais construits grâce à un ensemble de règles défini au sein d'un lexique. Ces règles permettent d'associer un concept (la conclusion de la règle) à une combinaison particulière de concepts pré-existants ou de termes (la prémisse de la règle) si certaines contraintes imposées sur les éléments des structures impliquées et sur leurs relations sont respectées. Le lexique est élaboré manuellement et expertisé. Un aperçu quantitatif de son contenu est proposé en fin de section dans la table 4.2.

Les concepts biologiques d'intérêt sont découverts à trois niveaux distincts d'analyse des *structures prédicat-arguments* grâce à trois ensembles indépendants de règles.

Les trois niveaux d'agrégation des concepts biologiques à partir des *structures prédicat-arguments* correspondent à trois stades d'abstraction successifs :

1. les concepts sont tout d'abord identifiés au sein de portions restreintes de texte (des mots ou des courtes séquences de mots) à l'intérieur de chaque élément des structures. C'est ce que nous appelons la lexicalisation. A cette étape, il n'existe encore aucun concept préalablement annoté dans les structures. Par exemple, "mRNA" et "gene transcript" se réfèrent tous deux au même concept *produit de la transcription*. L'élément de la structure correspondant à ces termes est désormais associé au concept. Certaines difficultés liées à la nominalisation sont prises en considération dès cette étape.
2. Dans un deuxième temps, les concepts existants vont être agrégés au sein d'une même structure grâce aux différentes combinaisons des éléments qui les contiennent et ainsi permettre de mettre à jour de nouveaux concepts. Par exemple, les concepts *locus de promoteur de gène*, retrouvé dans un élément ACTÉ, et *interaction moléculaire*, représentant une ACTION, donnent ensemble naissance à un nouveau concept intitulé *régulation du processus de transcription*. Les difficultés relatives à l'**alternance d'arguments internes** des verbes sont envisagées à cette étape.

3. Finalement, la découverte de certains concepts nécessite de mettre en relation les concepts préalablement identifiés, non plus au sein d'une structure unique, mais entre plusieurs structures. Par exemple, si une structure correspond au concept *production de protéine* et est reliée à une autre structure représentant le concept *activation de gène* par un lien exprimant la conséquence, un nouveau concept *activation de gène par protéine* est généré. Nous nous positionnons ici dans un contexte proche de la logique des propositions, au niveau de la phrase.

Il est à noter que les concepts retrouvés ne sont pas dépendants du niveau de l'analyse. Certains concepts sont communs à plusieurs niveaux d'analyse. Par exemple, le concept *activation d'une protéine par une protéine* peut être indifféremment retrouvé à partir des trois niveaux de l'analyse des concepts.

- Le concept est identifié à partir de l'expression "TCR IL2-mediated activation" grâce aux règles de niveau 1.
- Au niveau 2 dans l'expression "TCR activation takes place upon IL2 production".
- Et finalement au niveau 3 à partir de l'expression "TCR is activated when IL2 is produced".

De plus, plusieurs concepts peuvent être associés à une même portion des structures. Une même expression peut posséder différentes significations que seul le contexte de la phrase peut préciser. Le processus d'agrégation se charge alors d'éliminer certains concepts associés à une même expression à la faveur d'autres.

La liste des règles de conceptualisation a été établie manuellement à l'aide du **Metathesaurus UMLS**, des ressources internes de **LGP**, des données issues du dictionnaire de l'anglais courant référencé dans le paragraphe 3.2.3.2, du corpus de test présenté dans la section 3.2.4.2 et à la lecture de plus de 100 résumés **Medline** et d'une dizaine d'articles complets provenant de diverses sources.

Un effort particulier a été produit afin de prendre en considération la sémantique associée aux protocoles et aux expériences biologiques de part la natures des documents étudiés. Par exemple, dans la phrase "hTERT mRNA expression was found to increase in PPARγ knockout cells", le terme "knockout" permet de préciser que le gène PPARγ est soit absent soit rendu inopérant du génome de la cellule. Le traitement de l'information de la phrase, passée par le crible des différentes règles développées, nous permet au final d'établir que PPARγ réprime l'expression de hTERT.

4.3.2.1 Niveau 1 : lexicalisation

En avant-propos de ce paragraphe, il est important de souligner que l'approche de conceptualisation des structures que nous avons adopté n'est pas un travail à proprement parlé de lexique sémantique, notre vision du problème est très pragmatique et se limite uniquement à manipuler le lexique retrouvé dans un domaine de spécialité très restreint.

Nous nous efforçons ici de catégoriser le vocabulaire rencontré dans le cadre de la régulation de gènes. Cette catégorisation est elle même très simplifiée, un concept correspond, à ce niveau de l'analyse, à tout ou une portion des synonymes, hyperonymes et hyponymes d'un même champ lexical. Les antonymes sont eux généralement regroupés au sein de concepts indépendants.

L'ensemble de règles mis au point nous permet dans une premier temps de mettre en correspondance les termes de la phrase avec des concepts, puis, dans un deuxième temps, de réunir certains arrangements de ces concepts, découverts dans un même élément d'une structure, au sein d'un méta-concept. Les *POS* accompagnant les termes peuvent être utilisés afin de s'assurer de la validité d'une règle le cas échéant. D'autre part, la séquence d'apparition de plusieurs concepts au sein de l'élément peut être aussi un critère utilisé lors de la fusion de plusieurs concepts en un méta-concept. Par exemple, les deux expressions "T cell IL-2-dependant activation" et "IL-2 T cell-dependant activation" sont constituées des mêmes sous-concepts *cellule*, *protéine* et *activation*, acquis directement à partir de la terminologie. Néanmoins le méta-concept représenté par la mise en commun de ces concepts est différent dans les deux expressions et correspond dans le premier cas à *activation d'une cellule par une protéine* et de l'autre à *activation d'une protéine par une cellule*.

Nous avons isolé du lexique trois grandes catégories de termes :

- d'une part ceux qui décrivent ou qui dérivent d'une activité physique ou mentale. Ces termes peuvent être indépendamment retrouvés sous la forme de noms, de verbes ou d'adjectifs. Parmi cette grande catégorie nous pouvons dégager deux sous-catégories :
  - tout d'abord les notions non-spécifiques à la biologie comme celles liées au *changement* (par exemple, "to alter", "removal"), au *transport* (par exemple, "transmited", "to carry", "to pull", "set"), à la *localisation* (par exemple, "location"), à la *possession* (par exemple, "to provide", "possessing"), etc mais aussi les classes

de termes qui ne représentent pas une activité physique et qui sont spécifiques d'une interaction avec un être animé, telles que, par exemple, la notion que nous avons libellée *constatation* (par exemple, "to evaluate", "proving"). Ces notions sont utilisées uniquement lors de l'agrégation, en conjonction avec celles liées au vocabulaire de la biologie afin de former de nouveaux concepts.

- D'autre part, nous avons isolé du lexique les notions qui sont cette fois soit spécifiques ou qui ont une signification particulière dans le domaine de la biologie moléculaire. Les sous-catégories répertoriées ne représentent plus cette fois des classes abstraites mais des notions propres à la régulation de gènes. Par exemple, la *liaison chimique* ("binding", "to interact", "linking'), la *régulation de fonction* (par exemple, "regulated", "interaction', "to control", "alteration"), l'*activation de fonction* (par exemple, "stimulated", "overexpression", "to upregulate", "enhancement"), etc.

Chacune de ces notions peut donner naissance à des concepts distincts selon que le terme représente à proprement parlé l'action, est un effecteur de l'action ou une entité sur laquelle porte l'action. Des indices liés au *POS* (par exemple, "to activate", qui est un verbe, va décrire l'action et donc appartenir au concept *activation*), à la morphologie (par exemple, "activation" appartient au concept *activation* alors que "activator" est lui inclus dans le concept *agent d'activation*) ou encore grâce à l'agrégation de concepts dans le même élément de la structure (par exemple, "activation target", qui est composé de deux concepts distincts que sont *activation* et *cible*, permet de générer un méta-concept *cible d'activation*) permettent de préciser le concept du terme à partir de la notion à laquelle il appartient.

- D'autre part, nous avons répertorié des concepts propres aux êtres inanimés (les objets) et qui ne rentrent pas dans la première catégorie décrite. Sont uniquement pris en considération les termes spécifiquement utilisés en biologie moléculaire. L'ensemble des ENs rentre dans cette catégorie de concepts ainsi que les termes qui décrivent une classe d'objet biologique référencée (par exemple, "gene", "promoter") ou générique (par exemple, "sequence", "molecule", "domain", "mutant").
- Finalement, une dernière classe de concepts permet de prendre en compte les termes qui servent uniquement à structurer le discours. Ce sont des concepts abstraits qui rendent notamment en compte des opérations logiques qui organisent la phrase telles que la *cause* (par exemple, "to originate", "agent", "source", "to induce") et la *conséquence* (par exemple, "consequence", "outcome", "effect", "to result").

Il est à noter que de nombreux termes peuvent appartenir simultanément à plusieurs concepts.

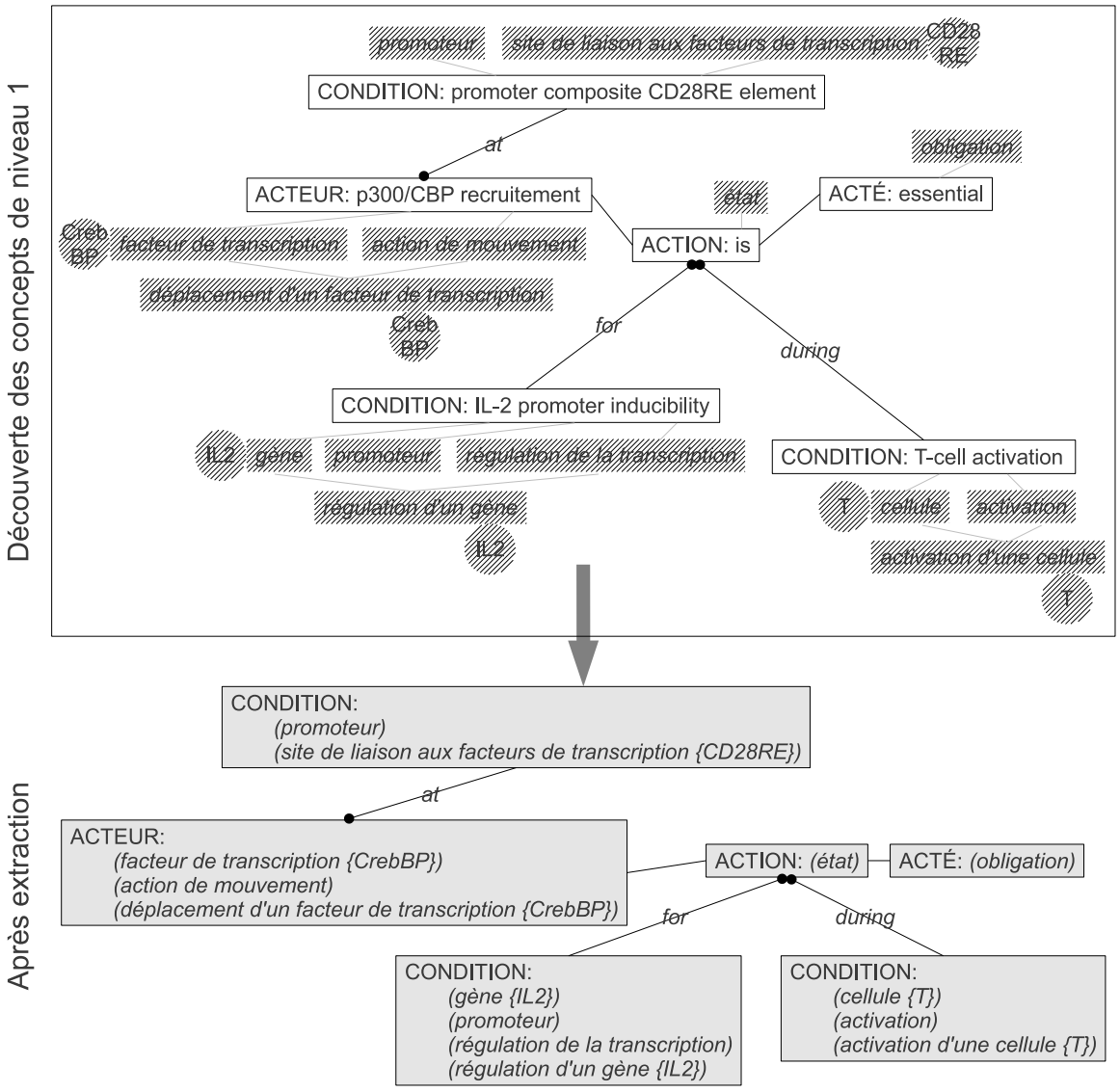
Un exemple du processus de conceptualisation à partir de règles de niveau 1 est proposée dans la figure 4.7. En haut de la figure est présentée la *structure prédicat-arguments* correspondant à la phrase "Recruitment of p300 at the composite CD28RE-TRE element of the promoter is essential for IL-2 promoter inducibility during T-cell activation". Certaines portions de texte ne correspondent à aucun concept et ne sont donc pas utilisables lors des conceptualisations ultérieures. La figure du bas correspond à la structure des concepts découverts. Chaque concept est signalé entre parenthèses et les ENs qui font partie d'un concept particulier sont notées entre crochets.

Nominalisation Le traitement des nominalisations verbales est une difficulté d'ordre syntaxique que nous nous efforçons de prendre en considération à cette étape. Les documents techniques, de part leurs style concis, sont riches en nominalisations. Certaines de ces nominalisations représentent une forme syntaxique alternative d'une *structure prédicat-arguments* autonome (toutes jusqu'à présent exclusivement déduites de la présence de groupes verbaux) : les groupes nominaux formés à partir d'une nominalisation verbale sont équivalents à des *structures prédicat-arguments* autonomes si leur *tête* convoit la notion d'ACTION.

L'inclusion du processus de résolution des nominalisations à l'étape de création des *structures prédicat-arguments* génériques s'est révélée être une stratégie inadéquate. L'interprétation des formes nominalisées n'est pas une tâche simple car d'une part les liens sémantiques qui unissent les membres du nom composé sont implicites et d'autre part parce que celle ci est intimement liée à l'usage et au contexte [Lap02]. Nous avons décidé de traiter le problème des nominalisations à l'étape de conceptualisation, plus à même de rendre compte des relations spécifiquement retrouvées dans les documents de biologie moléculaires entre les noms d'un groupe nominal issu du phénomène de nominalisation et non grâce à des règles purement syntaxiques. Nous avons développé un ensemble de règles qui nous permettent d'associer les mêmes concepts à la fois aux structures et aux éléments nominalisés s'il portent la même sémantique. La figure 4.8 illustre la découverte du même concept à partir d'une nominalisation verbale et d'une *structure prédicat-arguments* équivalente.

Les règles spécifiques à la gestion des nominalisation permettent de prendre en considération les appositions des ENs et concepts d'intérêt par rapport aux verbes nominalisés. En règle générale les noms situés devant le verbe nominalisés correspondent, par analogie, à des emplacements de type ACTÉ (par exemple, "JAK3 **activation**") alors que les noms placés derrière se réfèrent à des emplacements ACTEURs (par exemple, "JAK3 **ac-**

Fig. 4.7 – Exemple de conceptualisation à partir de règles de niveau 1



Les rectangles hachurés correspondent aux concepts découverts soit à partir de portions particulières du texte des éléments de la structure, et signalées grâce à de fines lignes grisées, soit grâce à des arrangements entre concepts existants. Les cercles hachurés sont les ENs qui font partie d'un concept identifié et par extension spécifient les instances de ces concepts.

Fig. 4.8 – Exemple de conceptualisation d'une nominalisation verbale (à gauche) et d'une structure prédicat-arguments sémantiquement équivalente (à droite).

tivation process"). Les règles développées sont presque exclusivement centrées autour des formes nominales de verbes décrivant une action physique (voir ci-dessus).

Verbes adjectivisés De la même manière que pour la nominalisation, certains verbes remplissant la fonction d'adjectif dans la phrase sont équivalents à des éléments ACTION de structures et font l'objet de règles particulières. Par exemple, les phrases "The experiment leads to an activated IL2", "The experiment leads to IL2 activation" et "The experiment results in activating IL2" sont représentées par des concepts identiques.

4.3.2.2 Niveau 2 : Conceptualisation des structures prédicat-arguments

A cette étape, nous cherchons à identifier les combinaisons significatives de concepts contenus dans les éléments d'une même *structure prédicat-arguments*. Les règles de conceptualisation utilisent pour uniques contraintes les concepts détectés à l'étape précédente de lexicalisation et les rôles (ACTEUR, ACTÉ, ACTION et CONDITION) des éléments de la structure concernée ainsi que les différents liens (prépositions et autres) qui unissent une CONDITION à un autre élément.

A ce niveau de l'analyse, deux phases distinctes sont nécessaires afin de dégager des concepts dérivés des structures.

1. Dans un premier temps, un ensemble de règles est spécifiquement préparé afin de découvrir des méta-concepts couvrant tout ou partie de la structure.
   - Soit les concepts sont issus de combinaisons particulières d'éléments satellites de la structure. Les élements sont souvent très peu nombreux et tous reliés directement entre eux. Par exemple, un élément ACTEUR ou ACTÉ représenté par le concept *activation d'un gène* et connecté à une CONDITION *cellule* par un lien tel que 'within' permet de générer un méta-concept *activation d'un gène dans une cellule* par agrégation des deux éléments.
   - Soit les méta-concepts sont identifiés à partir du concept convoyé par l'ACTION et des combinaisons des concepts contenus au sein des éléments ACTEUR, ACTÉ et CONDITIONs connectés. Les phénomènes d'**alternance d'arguments internes** des verbes de l'ACTION sont une difficulté que nous nous efforçons de traiter grâce aux règles développées (voir ci-dessous). Nous distinguons ici deux types de règles selon la sémantique de l'ACTION de la structure. Dans un premier cas, l'ACTION permet de structurer directement une relation d'intérêt dans le cadre de la régulation de gènes. Par exemple, un facteur de transcription *se lie*

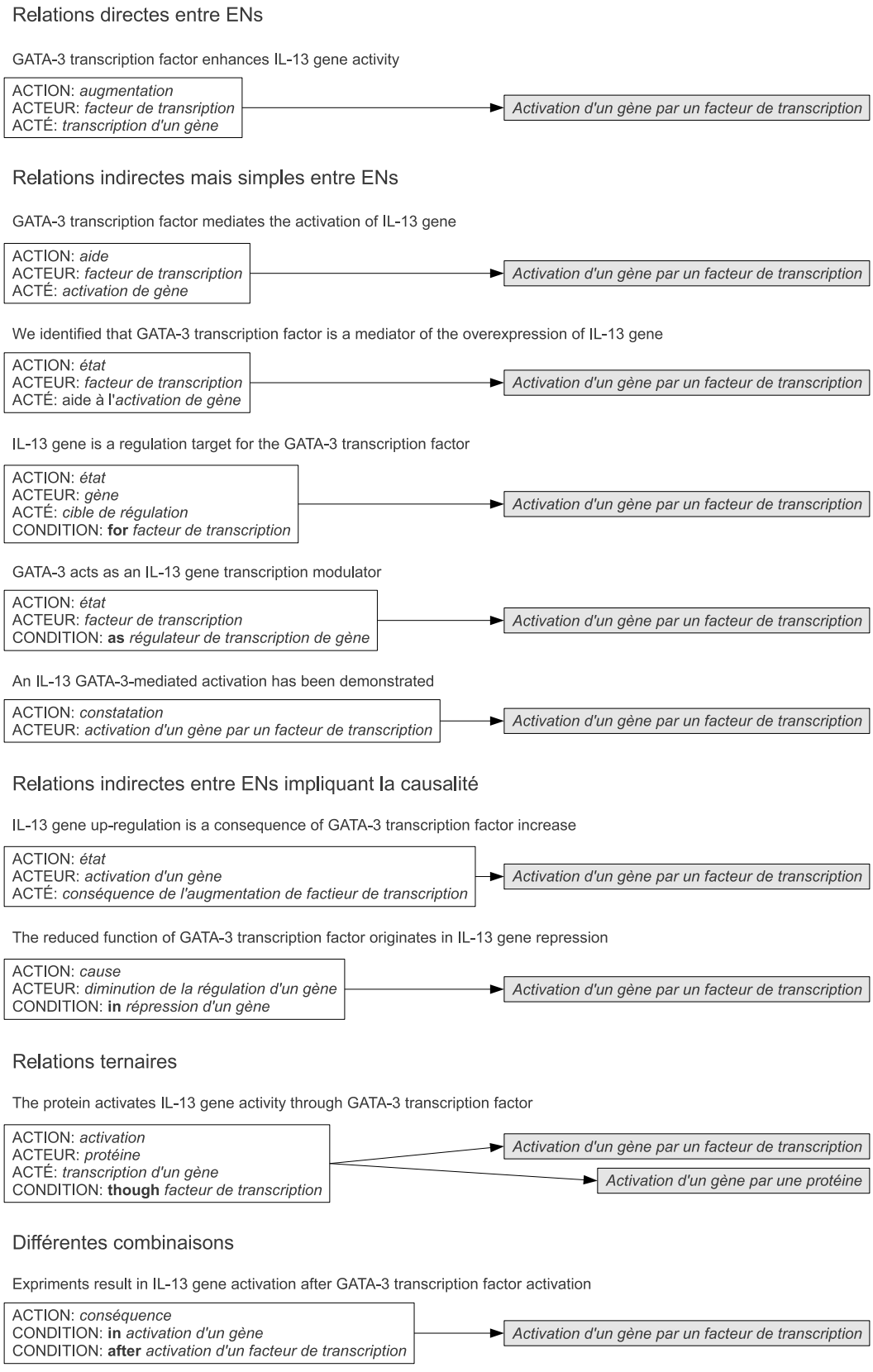
à un site de liaison aux facteur de transcription ou une protéine *active* l'expression d'un gène. Dans un deuxième cas, l'ACTION sert uniquement à connecter indirectement deux informations (ou plus) relatives à la régulations des gènes. L'ACTION peut préciser une association particulière telle que

- l'appartenance à une catégorie (par exemple, si un facteur de transcription *est identifié comme étant* l'activateur d'un gène alors le facteur de transcription *active* le gène),
- l'observation d'un mécanisme simultané (par exemple, si un facteur de transcription *contrarie* l'activation d'un gène, la répression d'un gène est conjointement observée à la présence d'un facteur de transcription et implicitement il est indiqué que le facteur de transcription réprime le gène)
- ou une forme de dépendance telle que la relation de cause et de conséquence (par exemple, si l'activation d'une protéine *permet* l'activation d'un gène, la répression d'un gène est observé après la répression d'une protéine ou la répression d'un gène *est due à* la répression d'une protéine alors la protéine *active* le gène).

2. Dans un deuxième temps, nous nous efforçons à mettre en relation les concepts précédemment découverts à cette étape. Les règles mises au point nous permettent d'agréger deux concepts ou plus lorsqu'une EN est commune aux concepts manipulés. Si une même EN permet de relier deux concepts selon les conditions établies par les règles alors ces derniers sont agrégés. Par exemple, dans la phrase "The CD28-responsive enhancer in the Human Interleukin-2 receptor alpha-chain locus is activated by CREB/ATF and CREB/ATF has been previously detected in lymphocytes", le concept représenté par la première portion de la phrase est *l'interaction d'un facteur de transcription avec un site de liaison au facteurs de transcription* alors que le deuxième concept est *production d'un facteur de transcription par une cellule*. le *facteur de transcription* instancié dans les deux concepts étant identique (CREB/ATF), nous considérons que ces deux concepts peuvent être agrégé en un seul concept *l'interaction d'un facteur de transcription avec un site de liaison au facteurs de transcription dans une cellule*. Certaines règles permettent de manipuler plus de deux concepts à la fois. Dans ce cas, chaque concept doit être au minimum relier à un autre concept par une EN commune. En l'état actuel des règles, seules les ENs sont utilisées afin de connecter et d'agréger les concepts isolés entre eux.

Un panel de structures représentant un même concept et affichant des constructions différentes est exposé dans la figure 4.9.

Fig. 4.9 – Exemples de structures prédicat-arguments représentant le concept d'activation d'un gène par un facteur de transcription et pris en charge par des règles conceptuelles de niveau 2



Un exemple de conceptualisation à partir de règles de niveau 2 est proposé dans le figure 4.10. En haut de la figure est présentée la *structure prédicat-arguments* préalablement conceptualisée par les règles lexicales et correspondant à la phrase utilisée dans l'exemple précédent de conceptualisation de niveau 1. Des concepts couvrant des portions éparses de la structure sont tout d'abord identifiés à partir des concepts lexicaux présents, puis combinés dans le cas où des ENs sont partagées afin de former des concepts additionnels. La partie du bas de la figure présente l'ensemble des concepts extraits à cette étape.

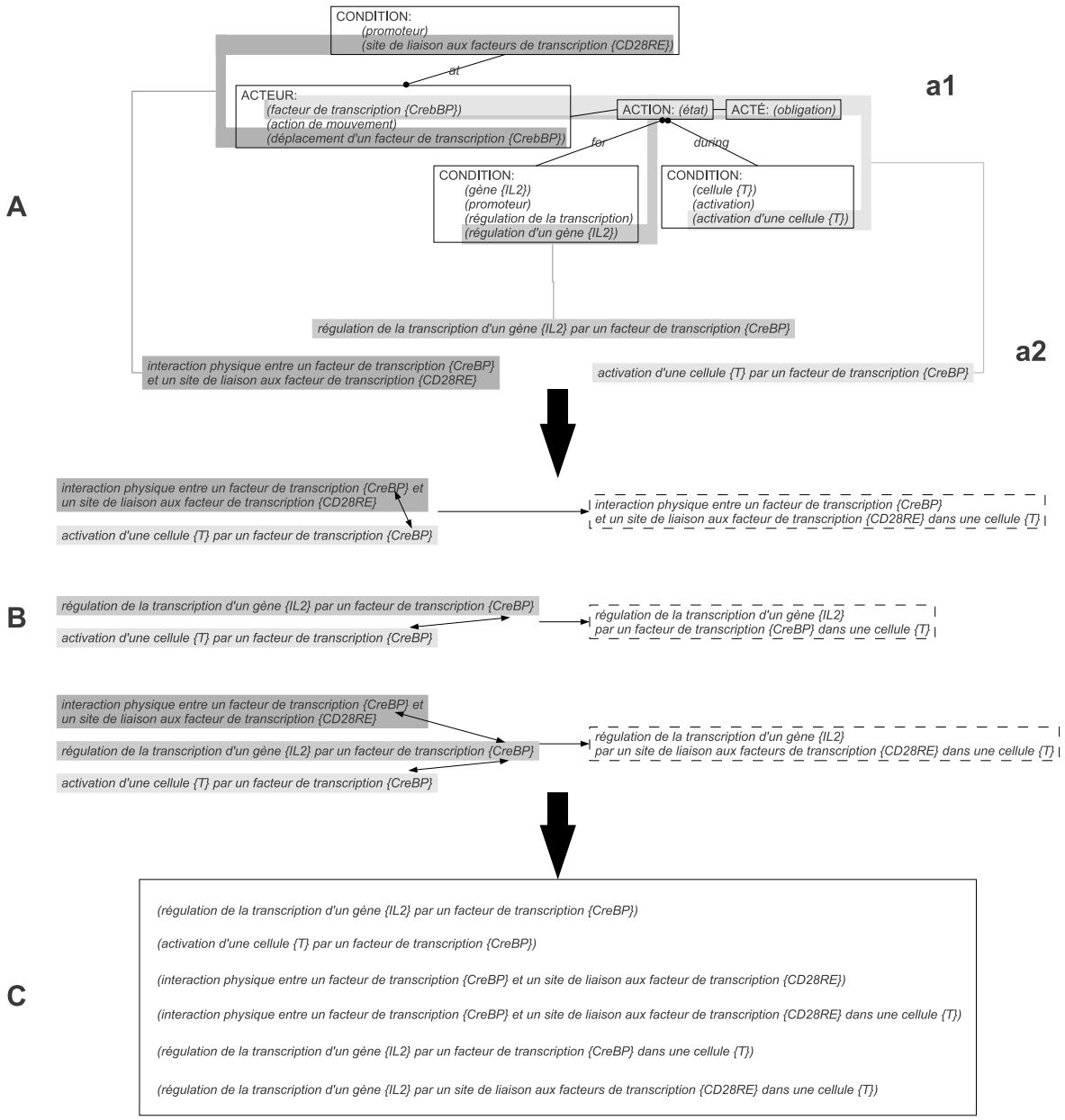
L'alternance d'arguments internes. A cette étape sont pris en considération les difficultés liées au phénomène d'**alternance d'arguments internes**. En d'autres termes, la difficulté consiste à identifier un même concept à partir des différentes combinaisons d'une ACTION et des éléments qui en dépendent.

Les règles développées permettent de rendre compte de certaines formes communes d'alternances que sont :

- La diversité des CONDITIONs dépendantes d'une ACTION. Les différentes prépositions qui introduisent une CONDITION peuvent tout aussi bien être des marqueurs redondants d'une même relation (par exemple, "IL2 regulates the TCR from T cells" et "IL2 regulates the TCR in T cells") ou au contraire spécifier deux informations distinctes (par exemple, "IL2 binds the TCR *through* its alpha chain" et "IL2 binds the TCR *for* an extended period").
- L'inter-inchangeabilité d'un ACTEUR ou d'un OBJET avec une CONDITION et le basculement de la transitivité d'une ACTION. Certaines CONDITIONs possèdent un rôle déguisé d'ACTEUR ou d'ACTÉ d'une structure. Par exemple, "The TCR is activated upon IL2 stimulation" et "The TCR is activated within T cells", dans le premier cas la CONDITION introduite par "upon" est l'ACTEUR de la structure alors que dans le deuxième cas la CONDITION introduite par "within" ne l'est pas. La transitivité d'un verbe d'une ACTION peut même être changée selon la fonction effective d'une CONDITION au sein de la structure. Par exemple, "TCR activation modulates upon IL2" et "IL2 modulates TCR activation" représentent la même information, néanmoins, et bien que la relation animant "TCR activation" et "IL2" soit orientée, l'ACTEUR d'une structure peut être la CONDTION de l'autre.

Nous ne nous positionnons pas dans une approche fondamentale de compréhension des phénomènes d'*alternances des arguments internes des verbes* et de leur sémantique. Notre vision est encore une fois pragmatique et centrée sur les exemples rencontrés dans le domaine de la biologie moléculaire. Certaines approximations grossières ont été réalisées

Fig. 4.10 – Exemple de conceptualisation à partir de règles de niveau 2



[A] Agrégation des concepts de niveau 1 issus des éléments de la *structure prédicat-arguments* : les concepts des éléments de la structure mis en valeur par une même nuance de gris [a1] sont agrégés afin de former un méta-concept [a2].
[B] Agrégation des concepts grâce à la présence d'ENs communes : un ensemble de règles distinct est utilisé afin de regrouper les concepts découverts à l'étape [A] en méta-concepts (représentés par des rectangles hachurés) s'ils partagent une EN dite de pivot.
[C] Liste de l'ensemble des concepts de niveau 2 découverts.

dans un souci de simplification de conception des règles d'extraction. Ainsi, nous avons décidé de développer les règles spécifiques à ce phénomène non pas autour du verbe de l'élément ACTION mais à l'aide du concept contenu dans ce même élément et ce dans un souci de simplification. Cette décision est critiquable car le phénomène d'*alternance* est très souvent spécifique à un verbe donné et non à une ensemble de verbes partageant un même concept [Pin89].

4.3.2.3 Niveau 3 : Conceptualisation des relations entre structures

A ce stade, nous cherchons à mettre en relations les concepts issus des différentes structures d'une même phrase.

Les concepts manipulés sont néanmoins identiques à ceux du niveau précédent. Les règles du niveau 2 qui permettent de prendre en considération les nominalisations et les adjectivisations verbales utilisent et génèrent les mêmes concepts que les règles de ce niveau ci. Par exemple, le concept *activation d'un gène par une protéine* peut être indifféremment retrouvé à partir de l'expression "CD158a/b expression is up-regulated after IL-2 is incubated" grâce à une règle de niveau 3 ou dans la phrase "CD158a/b expression is up-regulated by IL-2" par le biais d'une règle de niveau 2.

Deux classes de règles s'opposent ici encore :

- la première catégorie de règles est spécifique au cas où deux structures sont unies grâce à un lien explicite. Une préposition, ou une conjonction de subordination par exemple, relie les deux structures, une deux structures est utilisée en tant que CONDITION de l'autre.
- La deuxième catégorie de règles est caractérisée par l'absence de lien explicite entre deux structures. Comme pour les règles de niveau 2, la présence d'ENs communes entre les concepts ciblés par les règles est une condition nécessaire à leur agrégation.

Tab. 4.2 – Nombre de règles de conceptualisation utilisées

	Niveau 1	Niveau 2	Niveau 3	Total
Nombre de règles	1222	820	703	2745

De nombreuses règles partagent à la fois un même ensemble de concepts en prémisse ainsi que des concepts conclusions identiques entre les trois niveaux. Environ deux-tiers (estimation grossière) des règles du niveau 3 sont ainsi redondantes avec celles du niveau 2.
Ceci est principalement du à la gestion concurrentielle du phénomène de nominalisation sur l'ensemble des niveaux.

4.3.3 Les concepts à l'échelle de la phrase et du document

L'ensemble des concepts identifié grâce aux règles de niveau 1, 2 et 3 représentent l'information extraite à l'échelle de la phrase. Seuls les concepts référencés dans la liste 4.6 sont finalement conservés car ils représentent des données de notre modèle de la régulation de gènes. Dans le cas où plusieurs instances de concepts identifiés sont unis par une relation de spécialisation/généralisation, nous ne gardons que le concept le plus précis et omettons les autres. Par exemple, si dans une même phrase les deux concepts *activation d'un gène {MMP9} par une protéine {TNFα}* et *activation d'un gène {MMP9} par une protéine {TNFα} dans une cellule {myocyte}* sont retrouvés simultanément, seul le deuxième concept est conservé.

Tous ces concepts d'intérêt sont ensuite collectés et structurés dans notre base de données.

A cette étape apparait clairement la difficulté relative à la portée de ces concepts une fois qu'ils ont été identifiés. Le plus grand problème rencontré est alors la présence simultanée de nombreuses données contradictoires. Ces données, dans notre modèle de la régulation de gènes, correspondent à des concepts impliquant les notions opposées d'activation et de répressions tout en manipulant les mêmes ENs. Il est important de mettre en perspective ces différents concepts par rapport au contexte de leur apparition et de les replacer au sein du discours.

La compréhension des mécanismes d'enchainement du discours et de sa structuration est hors des propos de ce paragraphe et reste une des problématiques les plus complexes du TALN [PW05]. Nous ne cherchons pas non plus à confronter des données fragmentaires afin de les compléter, car sorti de l'échelle de la phrase, les liens logiques et d'organisation du discours ne sont pas accessibles en l'état actuel de la stratégie employée.

Afin de prendre en considération l'apparition d'informations conflictuelles, nous nous positionnons à l'échelle du document. Nous considérons qu'un article scientifique constitue une unité de discours standardisée. Les auteurs proposent en très grande majorité une thèse centrale qu'ils suivent et développent tout au long de leur document. Deux classes principales de données contradictoires à la thèse principale exposée peuvent néanmoins être référencées dans un même document : d'une part la mise en correspondance de leurs résultats avec d'anciennes données sur le sujet (souvent en provenance d'un ou de plusieurs groupes de recherche différents) et d'autre part la présentation de résultats opposés étayant leur thèse car encadrés dans un contexte expérimental extrêmement spécifique qui n'est pas détectable par nos règles de conceptualisation.

Nous partons de l'hypothèse que les informations relatives à la thèse du document sont répétées de multiples fois et surpassent en nombre les occurrences potentielles des données allant à leur encontre. Empiriquement, nous avons établi un certain nombre de conditions requises afin d'éliminer les instances de concepts contradictoires :

- Les multiples occurrences marquées comme étant incertaine ou hypothétiques (voir la section 4.2.2.3) d'une instance d'un concept sont supprimées s'il existe au moins une instance considérée comme 'certaine' d'un concept opposé.
- Si le nombre des instances d'un concept est inférieur a celui du concept opposé au sein de la section Conclusion et du Résumé (les deux sections mettant explicitement en avant la thèse des auteurs) alors tous les occurrences de l'instance de ce premier concepts sont éliminées dans l'ensemble du document.
- Dans tous les autres cas, d'un concept donné et de son opposé, seule l'instance majoritaire est conservée.

La résolution des contradictions entre deux documents n'est pas gérée et ces cas restent donc pour le moment non résolus.

4.4 Évaluation des performances du système d'identification des relations entre ENs et de l'ensemble des méthodes

Cette étude propose d'évaluer les performances du processus d'identification des relations entre ENs dans le contexte d'un processus de FdT globale à partir de publications

scientifiques complètes. L'impact des méthodes d'EEN et d'IEN sur les résultats obtenus est pris en considération. Les ENs et les relations sélectionnées sont présentées dans la section 4.3.1.

4.4.1 Jeu de données

A notre connaissance, aucun jeu de test étalon pré-existant ne correspond spécifiquement à nos besoins. Nous n'avons pas réutilisé et adapté le corpus de test des performances du système d'IEN (voir la section 3.2.4.2) pour cette tâche car l'ensemble des données déjà expertisées a servi de base à la création de notre corpus d'entrainement pour le développement de règles de conceptualisation.

Nous devons d'abord différencier les classes biologiques des ENs manipulées dans notre modèle de la régulation de gènes, puis les interactions entre ces ENs doivent être spécifiées et orientées, finalement, la nomenclature doit être spécifique à l'espèce humaine. A cette fin, nous avons préparé un ensemble de 25 publications complètes (et non juste les résumés), annotées manuellement. Ce corpus contient 3352 phrases et a été élaboré à partir de différentes sources. 15 articles ont été aléatoirement sélectionnés en utilisant les mots-clefs **MeSH** 'human', 'transcription' et 'promoteur' sur l'ensemble de la base bibliographique **BioMed Central**[4]. Parmi ces publications, 5 ne contiennent aucune information pertinente sur la régulation de gènes mais ont été néanmoins conservées afin d'ajouter du bruit dans le corpus de test. De plus, 10 documents ont été arbitrairement récupérés à partir des publications référencées dans la base de données **TRANSFAC**. Ils proposent tous des résultats expérimentaux en rapport direct avec la transcription de gènes chez l'humain et font montre d'un plus grand éventail de dates de publication que ceux en provenance de **BioMed Central**. La base de données bibliographique **BioMed Central** a été créée au début des années 2000 alors que **Transfac** propose des articles datés des années 90.

Sur ce jeu de données, nous n'avons annoté que l'information d'intérêt en relation directe avec notre modèle de la régulation de gènes tout en nous efforçant de ne pas surcharger le travail d'annotation, tâche extrêmement couteuse en temps. Les concepts de plus haut niveau correspondant à ceux présentés dans la section 4.3.1 (et reformulés dans le tableau 4.4) ont été annotés dans l'ensemble des documents. Seules les ENs présentes dans un concept d'intérêt sont elles aussi annotées. En revanche, celles ci ne sont pas

[4] http ://www.biomedcentral.com/

formellement identifiées à cette étape. La vérification de l'identité d'une EN extraite du texte et sa recherche au sein de nos dictionnaires est la tâche la plus couteuse de l'étape d'annotation et n'est pas réalisée. N'est retenu pour l'annotation d'une relation d'intérêt que la compatibilité des classes biologiques avec le concept. En l'état actuel de notre système d'identification des ENs, les homonymies à l'intérieur d'une même classe d'objets biologiques ne sont pas résolues aussi il n'est pas nécessaire lors de l'annotation de s'assurer de l'identité exacte d'une EN. Lors de la mesure des performances du système et dans le cas où un concept extrait par le système d'une phrase du jeu de données test est identique à un concept préalablement annoté dans cette même phrase, nous procédons néanmoins à deux vérifications distinctes sur les ENs extraites. D'une part, nous nous assurons que la localisation de l'EN extraite dans la phrase correspond à celle précisée dans l'annotation et d'autre part nous contrôlons que l'identifiant associé à l'EN extraite correspond bien à l'entité biologique dont parle l'auteur du document (d'autres entités homonymes peuvent être aussi rattachés à cet identifiant mais leur présence est alors ignorée). En contrepartie de cette décharge dans le travail d'annotation effectué, il nous est impossible de proposer la mesure des performances des méthodes d'identification des relations entre ENs seules sur l'ensemble de ce corpus de test. Pour ce faire, il nous faudrait annoter l'intégralité des ENs des textes.

Les sections des documents liées uniquement à la description des méthodes et des protocoles expérimentaux ont été rejetées pour les raisons évoquées en début de la section 2.5, c'est à dire à cause de leur contenu présumé pauvre en information. De la même manière, les références aux figures, tableaux et aux publications externes (les références bibliographiques) ont été automatiquement supprimées du texte. De plus, aucune anaphore ou co-référence n'a été prise en considération. Dans ce corpus, Nous avons estimé qu'approximativement 400 anaphores ou co-références sont impliquées dans des concepts d'intérêt et ont donc été omises de l'annotation. Les phrases informatives qui ne sont pas des affirmations fortes ne sont pas retenues dans le jeu de données. Ainsi, les constructions exprimant certaines formes d'incertitudes et d'hypothèses (voir la 4.2.2.3) ne sont pas annotées.

En complément, nous proposons aussi une mesure partielle des performances des méthodes d'identification des relations entre ENs uniquement. Chaque phrase contenant au minimum un concept annoté est isolée et réunie au sein d'un deuxième corpus de test. Seule la procédure de découverte de concepts est mise en œuvre à partir des ENs annotées sur ce second jeu de données de test.

4.4.2 Métriques

Afin d'évaluer les performances globales du système nous utilisons à la fois les mesures de précision et de rappel (voir le paragraphe 2.4.1.1). Les vrais positifs sont les instances d'un concept d'intérêt correctement identifiées par le système, cela implique que les ENs spécifiées dans ces concepts doivent être elles aussi identifiées avec justesse. Les faux positifs représentent les instances de concepts qui ont été incorrectement identifiées par le système en tant que vrais positifs. Finalement, les faux négatifs correspondent aux instances de concepts qui auraient du être identifiées par le système en tant que vrais positifs mais qui, par erreur, n'ont pas été détectées. Par exemple, une instance de concepts est à la fois comptabilisée en tant que faux positif et faux négatif lorsque le concept prédit est différent de celui précisé par l'annotation ou lorsqu'au minimum une EN impliquée est faussement identifiée (associée à une mauvaise classe biologique ou à un mauvais identifiant).

Nous utilisons ces métriques dans deux environnements de test distincts, un reflétant les performances brutes du système et l'autre plus proche des attentes d'un biologistes où nous ne nous positionnons plus à l'échelle de la phrase mais au niveau du document dans son intégralité. Dans le premier environnement d'évaluation du système, la précision et le rappel sont calculés pour chaque instance d'un concept rencontrée dans l'ensemble des documents, au complet. Dans le deuxième, nous procédons tout d'abord à la résolution des données contradictoires comme décrite dans la section 4.3.3, des concepts restants tous les doublons (c'est à dire les instances de concepts dupliquées dans un même document) sont omis lors du calcul des deux métriques. Dans ce cas, nous donnons la même importance à l'information considérée comme banale et à l'information dite 'exotique' au sein d'un même document. Dans notre jeu de données, l'information est très souvent reprise et répétée tout au long d'un document (voir la table 4.4) alors que d'autres concepts ne sont représentés qu'une seule et unique fois. Du point de vue d'un biologiste, la pertinence de l'information collectée n'est pas basée sur son ubiquité relative. Au contraire, les données rares sont habituellement considérées comme aussi informatives (parfois même plus) que les données sur-représentées.

Tab. 4.3 – Performances de l'identification des relations entre ENs sur le jeu de données test

Concept (relation d'intérêt entre ENs)	Nombre d'occurrences annotées	Rappel (partiel)[a]
P active la transcription de G	373	0.5
P régule la transcription de G	121	0.4
P réprime la transcription de G	58	0.35
T active la transcription de G	28	0.5
T régule la transcription de G	56	0.35
T réprime la transcription de G	2	1
T interagit avec S	222	0.3
S active la transcription de G	54	0.5
S réprime la transcription de G	10	0.9
[b]Un des concepts ci-dessus est localisé dans C	74	0.4
S appartient au promoteur de G	224	0.35
S est situé à BP sur le promoteur	113	0.35
S est composé de N	8	0.9
Total	1260	0.4

[a]Mesure calculée sur une portion du corpus de test correspondant aux phrases contenant des concepts annotés et non sur l'ensemble des textes.
[b]Dans un souci de simplification, les mesures relatives aux relations ternaires sont présentées globalement et non individuellement.

4.4.3 Résultats

La table 4.3 montre les résultats obtenus sur notre jeu de test réduit, basé sur l'utilisation des méthodes d'identification de concepts uniquement. La lettre P symbolise une instance d'une protéine ou d'un gène, G une instance d'un gène uniquement, T une instance d'un facteur de transcription, S une instance d'un site de liaison aux facteurs de transcription, C une instance d'une lignée ou d'un type cellulaire, BP une instance d'une position relative (mesurée en paires de bases) à au site de démarrage de la transcription d'un gène et N une instance d'une séquence d'acides nucléiques.

Le rappel partiel obtenu est approximé à 40%.

La table 4.4 montre les résultats obtenus sur notre jeu de test au complet, basés sur l'utilisation combinée de nos méthodes d'EEN et d'IEN et d'identification de concepts.

Nous rapportons une précision globale de 67% et un rappel de 23%. Si nous nous positionnons à l'échelle du document et non de la phrase et que nous ne nous prenons plus en considération le nombre d'occurrences des concepts à extraire au sein d'une même publication, la précision demeure presque inchangée (62%) alors que le rappel croit de

Tab. 4.4 – Résultats expérimentaux globaux sur le jeu de données test

Concept (relation d'intérêt entre ENs)	Nombre d'occurrences annotées	Vrais positifs	Faux positifs	Faux négatifs	Précision	Rappel
P active la transcription de G	373	97	50	276	0.66	0.26
P régule la transcription de G	121	23	10	98	0.7	0.19
P réprime la transcription de G	58	10	3	49	0.77	0.17
T active la transcription de G	28	8	3	21	0.72	0.28
T régule la transcription de G	56	21	7	103	0.76	0.17
T réprime la transcription de G	2	2	0	0	1	1
T interagit avec S	222	47	37	177	0.56	0.21
S active la transcription de G	54	11	3	44	0.78	0.2
S réprime la transcription de G	10	7	1	3	0.87	0.7
*Un des concepts ci-dessus est localisé dans C	74	16	8	60	0.66	0.21
S appartient au promoteur de G	224	60	26	162	0.7	0.27
S est situé à BP sur le promoteur	113	20	7	91	0.75	0.18
S est composé de N	8	8	0	1	1	0.87
Total	1260	330	155	1085	0.68	0.23
Concept (relation d'intérêt entre ENs)	Nombre d'instances annotées par document	Vrais positifs	Faux positifs	Faux négatifs	Précision à l'échelle du document	Rappel à l'échelle du document
P active la transcription de G	103	46	27	57	0.63	0.44
P régule la transcription de G	69	18	10	51	0.64	0.26
P réprime la transcription de G	18	6	3	12	0.66	0.33
T active la transcription de G	18	6	3	12	0.66	0.33
T régule la transcription de G	14	5	3	9	0.62	0.35
T réprime la transcription de G	2	2	0	2	1	1
T interagit avec S	126	34	36	92	0.48	0.27
S active la transcription de G	26	7	3	19	0.7	0.27
S réprime la transcription de G	6	2	1	4	0.66	0.33
*Un des concepts ci-dessus est localisé dans C	28	12	7	16	0.63	0.42
S appartient au promoteur de G	95	27	13	68	0.67	0.28
S est situé à BP sur le promoteur	39	14	5	25	0.73	0.38
S est composé de N	7	7	0	0	1	1
Total	520	186	111	367	0.62	0.34

*Dans un souci de simplification, les mesures relatives aux relations ternaires sont présentées globalement et non individuellement.

moitié (34%). Cette hausse du rappel est expliquée par le fait que l'information est en majorité non seulement répétée mais est encore reformulée dans la plupart des cas tout au long d'un document dans notre corpus de test. Les occasions de récupérer un concept unique par document sont alors plus nombreuses. Cependant, nous constatons que cette redondance est souvent restreinte aux concepts en relation directe avec le thème central de la publication. La légère diminution de la valeur de la précision est majoritairement une conséquence de la singularité des différents faux positifs. Comme discuté ci-dessous, un faux positif est souvent le résultat d'une erreur liée à un contexte textuelle spécifique et, en tant que tel, il est peu probable qu'une même erreur se reproduise sur l'ensemble du document. Nous n'avons pas comptabilisé, après annotation, le nombre de concepts en contradiction dans un document, néanmoins il semble que cela soit un phénomène peu représenté au sein de notre corpus de test. Il est important de rappeler qu'aucune phrase exprimant une incertitude n'a été retenue dans le corpus de test. Aussi, les structures marquées comme incertaines par le système ne sont pas comptabilisées dans les résultats présentés dans le tableau 4.4. Nous avons ainsi automatiquement éliminé environ 7% du nombre total des concepts identifiés sur l'ensemble du jeu de données.

Bien que la population des différents concepts annotés soit dissemblable, le système affiche un comportement homogène quelque soit le concept à extraire et ce avec une exactitude comparable. Ceci peut être principalement expliqué grâce à deux facteurs. D'une part, les méthodes mises au point et les ressources utilisées sont partagées pour tous les concepts. D'autre part, les difficultés rencontrées sont globales et communes aux différents concepts. Néanmoins, quelques spécificités peuvent être observées. Ce dernier point est principalement du au degré de polymorphisme de certains concepts. Par exemple, les relations qui impliquent les concepts liés à la *régulation* manifestent globalement un rappel plus faible que leurs homologues centrés autour des concepts connectés à l'*activation* ou la *répression*. La *régulation* est une relation plus générique que l'*activation* ou la *répression*.

Il est extrêmement difficile de comparer ces résultats avec d'autres préalablement publiés dans la littérature. Le travail de Saric *et al.* (voir la section 2.5) converge sur certains points avec le nôtre, soit la détection de relations entre des gènes et des protéines dans le contexte de l'expression de gènes. Cette partie de leur étude est grossièrement équivalente aux 6 premiers concepts présentés dans le tableau 4.4. Cependant, leur exploration des relations entre ENs n'est pas en relation avec la physiologie humaine et, tout comme nous l'avons fait, ils ont développé leur propre corpus de test centré autour de leurs besoins exclusifs. De plus, les auteurs ont fait le choix de ne mesurer les performances de leur

système que sur un sous-ensemble très restreint de leur jeu de données : la précision est calculée sur 90 relations, le rappel n'est pas mentionné.

Tab. 4.5 – Échantillon aléatoire de 60 causes d'erreurs expérimentales sur le jeu de données test

Type de la cause	Nombre d'occurrences
Contenu des dictionnaires d'ENs	18
Extraction ou identification de *structures prédicat-arguments*	16
Analyse syntaxique	13
Gestion des coordinations	5
Désambiguisation des ENs	8
Confusion gène-protéine	3
Confusion facteur de transcription-site de liaison	3
Divers	5

Afin d'estimer de recenser et de catégoriser les principales carences du système dans son ensemble, nous avons aléatoirement collecté 60 causes d'erreurs à partir du jeu de données test. Ces causes d'erreurs peuvent être tout aussi bien générer des faux positifs que des faux négatifs et sont présentées dans le tableau 4.5. Il est important de noter que ces chiffres ne tiennent pas compte du caractère cumulatif des causes pouvant conduire à une erreur. Certaines causes sont artificiellement dissimulées car elles surviennent lors de phases ultérieures du traitement des phrases. Les causes d'erreurs rencontrées chronologiquement lors du traitement des phrases sont tout d'abord liées au dictionnaires des ENs, puis à la désambigüisation des ENs, à l'analyse syntaxique et finalement à l'extraction des concepts d'intérêt. Il en résulte que les sources d'erreurs liées par exemple à l'analyse de la syntaxe sont minimisées alors que les chiffres obtenus pour les problèmes issus des dictionnaires sont relativement fidèles dans notre panel de causes d'erreurs.

1. Parmi les difficultés les plus représentées, les déficiences du contenu des dictionnaires des ENs tiennent une place prépondérante. Les causes sont identiques à celles discutées dans la section 3.1.4. Ce type d'erreurs conduit principalement à la présence de concepts faux négatifs (environ 75% des erreurs de ce type) et est soit la conséquence d'une absence globale de l'entité ou l'omission d'une forme variante spécifique dans les dictionnaires. Cette difficulté est généralement artificiellement propagée à l'intégralité d'un document lorsque l'EN non détectée est connectée au thème central de la publication. Il est aussi important de noter que la grande majorité des auteurs des articles de notre corpus de test conserve la même forme variante d'une EN dans l'ensemble du document. En conséquence, ceci se résume à une situa-

tion de 'tout ou rien' dans la détection des ENs. Nous pouvons soit correctement, et avec facilité, détecter chaque instance d'une EN spécifique dans une publication ou au contraire manquer toutes les occurrences si la forme variante est absente des dictionnaires. Les faux positifs sont en revanche moins fréquents (environ 25% des erreurs de ce type). Parmi les faux positifs référencés, la très grande majorité de ces erreurs est uniquement due à des problèmes de spécification des objets biologiques hyperonymes/hyponymes au sein de nos dictionnaires. Cette cause a déjà été signalée dans le paragraphe 3.2.4.2. Certaines entités, faussement identifiées, correspondent à une portion du nom de l'entité réelle du texte uniquement car cette dernière n'est pas enregistrée dans les dictionnaires. Ce phénomène semble être plus particulièrement spécifique aux facteurs de transcription et aux sites de liaison aux facteurs de transcription dans notre jeu de données test. Malgré les précautions prises lors du nettoyage manuel des dictionnaires, des noms de familles d'entités peuvent être utilisés en lieu et en place des noms des membres précis de ces mêmes familles. Par exemple, "NF-κB3" (aussi connu sous les noms "RelA" et "p65 NF-κB") est incorrectement identifié en tant que "NF-κB", le nom générique de la famille à laquelle l'entité appartient, dans notre corpus de test. De part la nature expérimentale des documents étudiés, de nombreuses constructions artificielles d'objets biologiques sont manipulées. Ces entités particulières partagent très souvent une portion du nom de l'objet biologique qui a permis leur construction. Néanmoins, ces deux entités sont par définition différentes. C'est par exemple le cas des protéines dites chimériques (de fusion). Elles peuvent comporter simultanément, dans le texte, les noms des protéines fusionnées comme dans le cas de "Lym-1/Interleukin 2", dans d'autres contextes le caractère '/' est utilisé en tant que conjonction de coordination 'ou' ou 'et'.

2. Une autre source importante d'erreurs d'extraction de concepts est liée à la définition des règles de conceptualisation dans le contexte de la régulation de gènes. Alors que l'exactitude des règles existantes semble acceptable, leur nombre et les combinaisons réalisables sont largement insuffisantes afin de couvrir toutes les représentations textuelles de l'information trouvée dans le jeu de données. Ce type d'erreurs est une source de faux négatifs.
   - La première insuffisance détectée dans cette catégorie d'erreurs est à mettre en correspondance avec l'absence de longues séquences de concepts inter-connectés et agrégeables. Les concepts des plus hauts niveaux d'abstraction peuvent être indirectement retrouvés grâce à la combinaison de concepts de plus bas niveaux, comme nous l'avons vu dans la section 4.3.2. Néanmoins, à cause de contraintes de

temps essentiellement, nous n'avons conçu que des règles d'agrégation de concepts inter-*structures prédicat-arguments* utilisant une EN en tant que 'point de pivot' entre les concepts de plus haut niveau à combiner. Ainsi, les relations en *cascade* sont généralement perdues. Par exemple, dans les expressions suivantes : "only the locus bearing a GATA-3 site activates IL-4 promoter activity" et "GATA-3 is critical in opening the locus that regulates IL-13 transcription", l'annotation révèle qu'un site de liaison aux facteurs de transcription "GATA-3" est capable d'activer la transcription de "IL-4" et que le facteur de transcription "GATA-3" régule la transcription de "IL-13". Le système n'a pas détecté ces relations qui sont pourtant des instances valides de concepts clefs dans notre modèle de la régulation de gènes. Il est à noter que dans les constructions avec 'to' et l'infinitif ou 'for' et un participe présent, la difficulté n'est généralement pas visible. Par exemple, dans la phrase "IL2 is a protein known to activate the transcription DCS", les règles de simplification de la syntaxe nous permettent d'associer directement "IL2" en tant qu'acteur de l'ACTION "activate" et donc de passer outre la difficulté. D'après l'étude portant sur notre échantillon d'erreurs, ces 'points de pivot' dans la phrase sont tous détenteurs d'une sémantique propre à la classe de l'objet biologique à laquelle ils sont rattachés. Ce sont soit des termes génériques qui peuvent être référencés dans notre lexique de désambigüisation (voir la section 3.2.3.1) (par exemple, "the protein", "the ligand") ou des noms de familles, groupes, sous-structures, sous-unités et complexes (par exemple, "Pro-inflammatory cytokines, including interleukin-1 beta (IL-1 beta) or tumour necrosis factor alpha (TNF alpha), upregulate c-Jun protein in neonatal rat ventricular myocytes"). Ces termes peuvent ainsi être comparés à des co-références. Les concepts résultant de leur manipulation ont été néanmoins annotés car les ENs auxquelles ils se réfèrent sont tous simultanément présents dans la phrase. Environ deux tiers des concepts qui n'ont pas été identifiés à cause de cette difficulté ne comportent qu'un seul 'point de pivot'.

- Le second type de problèmes détecté, et toujours connecté à la définition des règles de conceptualisation, est à mettre en relation avec les lacunes apparentes soit du vocabulaire manipulé soit des règles d'agrégation développées.
  - Fréquemment, le problème survient lorsque le texte provenant d'un élément unique d'une *structure prédicat-arguments* ne peut être converti en concepts. Par exemple, le terme "turn on" n'est pas détecté comme faisant partie du concept *activation*. Cette difficulté précise est spécifique à l'étape de lexicalisation (le niveau 1 de conceptualisation).

- Occasionnellement, les lacunes dans la construction des règles sont visibles lorsqu'un nouveau concept ne peut être inféré à partir des combinaisons des concepts existants à l'intérieur d'une même *structure prédicat-arguments*. Par exemple, la portion de la structure représentant la construction soulignée dans la phrase "Basal expression of cyclooxygenase-2 shares the ability of nuclear factor-interleukin 6 to be coordinately regulated by interleukin 1" n'est associé à aucun concept et l'information que "nuclear factor-interleukin 6" est aussi régulée par "interleukin 1" est perdue.
- Il est aussi intéressant de noter que les idiomes et les expressions figurées ne sont pas gérés en l'état actuel du système. Par exemple, la phrase "IL-5 is expressed when the GATA brake is removed" implique que "GATA" normalement réprime l'expresssion de "IL-5", néanmoins cette information ne peut être traitée car le système n'a aucune idée de ce qu'un "GATA brake" signifie.

3. Les problèmes issus de l'analyse de la syntaxe constituent le troisième type d'erreurs majeures détectées. Ils ont en général un impact bien plus importants sur les performances du système que les autres formes d'erreurs recencées ici, même si quantitativement cela ne parait pas évident au vue de notre échantillon.
   - Les erreurs de traitement syntaxique observées impliquent souvent la liaison d'un verbe au mauvais sujet ou au complément d'objet. Dans ce cas, toute l'information contenue dans la phrase est soit perdue soit mal interprétée. Cette forme d'erreur est majoritairement observable lorsque le sujet ou le complément d'objet est relativement distant du verbe, généralement séparés par une longue proposition. Le problème apparait aussi lorsque le sujet ou le complément d'objet sont partie intégrante d'une énumération. **LGP** montre de grandes difficultés à gérer les conjonctions de coordination correctement. Sur ce type de phrase, il apparait que **LGP** est obligé dans la plupart des situations d'avorter la première passe de l'analyse et de se rabattre sur une analyse syntaxique partielle utilisant les *liens nuls* (voir la section 4.5). Il en résulte la présence d'ilôts de structures qui ne sont pas résolues. Ce type d'erreurs peut tout aussi bien produire des concepts faux négatifs comme des concepts faux positifs.
   - A part les phrases non-verbales que **LGP** ne peut analyser, les phrases avec un contenu hautement technique tel que les résultats chiffrés des expérimentations réalisées par les auteurs du document sont elles aussi plus sensibles aux erreurs liées au traitement de la syntaxe. Dans la majorité des cas, l'analyse de la phrase est avortée sans qu'aucune portion de la structure ne soit résolue. ce type d'erreurs engendre la présence de concepts faux négatifs. Les phrases issues de la

section Résultats des publications sont les plus difficiles à analyser à cause de cette difficulté. Beaucoup de résultats, à l'origine présentés dans la section Méthodes, sont répétées dans cette portion du document. **LGP** est à l'origine un analyseur syntaxique généraliste sans aucune adaptation particulière à la syntaxe mathématique.

4. La dernière source la plus fréquente d'erreurs est liée au processus de désambigüisation des ENs.
   - Tout d'abord, nous constatons de nouveau (voir la section 3.2.4.2) que le contexte nécessaire à la désambigüisation de l'identité d'une EN à l'intérieur d'une phrase se révèle parfois insuffisant. Un contexte plus large, couvrant plus d'une phrase, est souvent requis.
   - Ensuite, les méthodes développées sont inadaptées aux situations complexes. La syntaxe de la phrase peut être trop difficile à cerner sans l'aide de l'analyseur syntaxique, laissant les éléments textuels clefs pour la désambigüisation détachés des ENs correspondantes.

   Le problème qui est surtout observé est relié à la difficulté de discerner les facteurs de transcription des sites de liaison aux facteurs de transcription d'une part et les protéines des gènes d'autre part. Un exemple de phrase présentant cette difficulté est "First, we show that IL-6 activates HSP72". Seul le contexte de la phrase, à l'échelle du paragraphe dans cet exemple particulier, nous apprend que "IL6" est sous la forme protéique et que "HSP72" est sous la forme gène. Ce type d'erreurs génère à la fois des concepts faux positifs et faux négatifs.

5. Les autres erreurs détectées sont en rapport avec le contexte de la validité des relations détectées et la gestion des hypothèses. Les indices textuels nécessaires afin de s'assurer de la validité des relations extraites par le système sont absents de la phrase. La majorité de ces erreurs ne peut être résolue que grâce à l'utilisation de données disponibles uniquement au niveau du discours.
   - Par exemple, les résultats d'expérience peuvent être décrits sans mentionner leur source. Certains sont obtenus par des moyens bioinformatiques et, en tant que tel, ne peuvent être considérés avec le même indice de fiabilité que des rapports d'expérimentations *in vivo* et *in vitro*. De la même manière, certains résultats doivent être rejetés si nous savons qu'ils sont spécifiques à la souris et non à l'humain.
   - D'autres phrases expriment des hypothèses de travail et en aucun cas ne proposent des concepts utilisables dans notre modèle. Par exemple, dans les phrases suivantes : "First, we will ascertain whether IL-3 activates eosinophils" et "In this

section, p65 NF-kB repression of the increase of PEPCK gene transcription is evaluated" l'annotation précise qu'aucun concept ne doit être extrait or le système identifie que la protéine "IL3" active les cellules "eosinophils" et que "NF-KB" réprime la transcription de "PEPCK". Ce type d'erreurs est responsable de la détection de concepts faux positifs majoritairement.

– Finalement, nous avons observé certaines sources d'erreurs rares qui sont imputables à l'ambigüité syntaxique de certaines propositions relatives et locutions prépositives (voir les paragraphes et). Ces difficultés ont la particularité d'être très ardues, même pour l'annotateur, et génèrent des contresens. Les différentes propositions peuvent être rattachées aux mauvaises éléments de la phrase lors de l'analyse syntaxique sans pour autant créer d'erreurs de syntaxe ou d'impossibilités sémantiques. Par exemple, dans la phrase "IL-12 most strongly enhanced the gene expression of IL-2Ralpha, c-myc, and pim-1 with IL-2", "IL-12" active-t-il l'expression de l'ensemble de ces gènes ou uniquement de "pim-1" grâce à l'aide de "IL-2" ? L'annotation révèle que la première solution est correcte alors que le système avait retenu la deuxième. La proposition soulignée avait été rattachée à l'ACTÉ "pim-1" dans la *définition* correspondante lors de l'analyse syntaxique et non à l'ensemble des ACTÉs de chaque *définition*. Ce type d'erreurs peut être responsable de la présence de concepts simultanément faux positifs et faux négatifs. Dans notre jeu de données nous n'avons observé que des cas de faux négatifs.

4.5 Résumé

Nous avons proposé dans cette section une méthode d'extraction et d'identification de relations entre entités biologiques ainsi que la découverte de certaines de leurs caractéristiques en réalisant une analyse dite profonde de la syntaxe de la phrase. Nous avons montré la faisabilité d'une telle tâche une fois que le vocabulaire spécifique au domaine de l'étude a été acquis. De plus, le système peut être facilement adapté afin de détecter des concepts relatifs à l'expression de gènes qui ne soient pas déjà inclus dans notre modèle de la transcription de gènes. Dans les approches basées sur l'analyse syntaxique profonde de la phrase, la tâche la plus pénible, soit la conception d'heuristiques dans le but d'acquérir la sémantique à partir de la syntaxe, a été minimisée. D'une part, nous avons défini deux sous-tâches indépendantes, c'est à dire l'acquisition de règles linguistiques génériques et la conception de règles d'extraction spécifiques au domaine de l'étude. D'autre part, une hiérarchie de concepts reliés à notre modèle de la régulation de gènes simplifie le

développement des règles d'extraction.

Nous réalisons dans un premier temps l'analyse profonde de la syntaxe des phrases grâce à **LGP** que nous avons adapté au domaine de la biologie moléculaire. Afin de s'affranchir des formes redondantes de représentation syntaxique de l'information, nous structurons le contenu des phrases au sein de *structures prédicat-arguments* génériques à l'aide de règles de simplification de la syntaxe dédiées à la résolution de difficultés communes aux textes écrits en langue anglaise. Ces difficultés incluent l'absorption des différences entre les formes passives et actives, les tournures personnelles et impersonnelles et les structures *en montée* et directes. De ces *structures prédicat-arguments*, l'information sémantique est alors extraite grâce à l'utilisation de règles de conceptualisation propres au domaine de la régulation de la transcription de gènes. Cette phase d'extraction nous permet, dans un premier temps, de prendre en considération des difficultés d'ordre syntaxique non encore gérées par les *structures prédicat-arguments* telles que la nominalisation et le procédé d'*alternance d'arguments internes des verbes* puis, dans un deuxième temps, de détecter des relations pertinentes entre deux ou trois ENs impliquées dans le processus de la régulation des gènes. Ces relations peuvent être de type direct ou indirect et manipuler des rapports logiques simples telles que la cause et la conséquence.

Finalement, les performances globales du système ont été évaluées à partir d'un corpus de textes complets en biologie moléculaire. Ces résultats prennent aussi en considération les méthodes de REN et sont donc un indice de la fiabilité du processus de FdT global sur des données réelles. Nous extrayons et identifions les paramètres de notre modèle de la régulation des gènes avec une précision globale de 62% et un rappel de 34%. Les principales difficultés rencontrées lors de l'analyse sont liées d'une part à la présence d'ENs inconnues et d'autre part à la complexité de la syntaxe et des concepts manipulés dans les textes de biologie moléculaire.

Conclusion et perspectives

Nous conclurons cette dissertation en discutant de la contribution et des limites du système développé. Puis nous proposerons quelques voies futures de recherche.

Conclusion

Notre travail de thèse a consisté en la conception, l'adaptation et la mise en œuvre de méthodes de FdT à une problématique de biologie moléculaire encore peu étudiée dans le cadre d'une approche d'extraction d'information systématique : l'extraction des réseaux de régulation de l'expression des gènes à l'échelle du noyau de la cellule et à l'échelle de l'organisme.

Notre approche est globale en ce sens que nous avons proposé un ensemble de méthodes capable de prendre en charge un document textuel numérique jusqu'à la structuration de l'information pertinente qu'il contient. Nous avons tout d'abord présenté une procédure simple d'extraction et d'identification d'ENs complexes utilisant des dictionnaires spécialisés. Puis nous avons proposé une procédure de détection des relations qui animent ces différentes ENs ainsi que la récupération de certaines de leurs caractéristiques basées sur une analyse dite profonde de la syntaxe de la phrase.

Nous nous sommes efforcés de ne pas lier les solutions mises en œuvre aux spécificités de notre domaine d'étude, la régulation de gènes, lorsque cela n'était pas nécessaire. Aussi les méthodes développées se révèle être relativement génériques et peuvent être utilisées dans le cadre de la résolution d'autres problématiques de biologie en FdT. Par exemple, d'autres ENs peuvent être détectées dans les textes en enrichissant les dictionnaires à partir de nouvelles sources terminologiques et en adaptant le lexique de désambigüisation en conséquence. Un autre exemple, précédemment discuté, consiste en l'indépendance

des étapes d'acquisition de la syntaxe et de génération des *structures prédicat-arguments* génériques de l'étape de conceptualisation, spécifique au domaine de l'étude.

Néanmoins, de nombreuses limites et contraintes dans notre méthodologie sont apparentes. Le modèle de réseau proposé dans la section 2.5 qui est à la base du travail présenté dans ce document représente une simplification extrême du processus de régulation de l'expression des gènes eucaryotes. L'absence de modélisation de données cinétiques en est un exemple flagrant. De plus, nous avons éludé de nombreuses difficultés techniques lors de l'élaboration des méthodes :

- Les dictionnaires d'ENs peuvent être très difficilement mis à jour car un effort manuel important de nettoyage et de validation de leur contenu est requis afin de leur assurer une fiabilité minimale. De plus, la structuration actuelle de nos dictionnaires ne nous permet pas de gérer des relations conceptuelles simples telles que la spécialisation/généralisation (par exemple, famille $\rightleftharpoons$ membre $\rightleftharpoons$ domaine ou sous-unité). Ainsi, il serait utile de ne plus stocker les références des ENs au sein de dictionnaires ou de lexiques mais de les intégrer dans des ontologies spécialisées. Nous pensons que cela pourrait augmenter les performances du système et diminuer à la fois le nombre de faux positifs et de faux négatifs.
- Lors de l'étape de REN, nous avons constaté que les techniques de désambigüisation développées se révèlent être insuffisantes. Nous imputons cette déficience au choix trop restreint des termes de désambigüisation dans notre lexique.
- Nous avons aussi éludé le problème des co-références et des anaphores. Leur résolution permettrait d'explorer l'ensemble des documents et non juste les portions où sont présentes les ENs explicites.
- Finalement, il apparait qu'un effort particulier doit être apporté dans la conception des règles de conceptualisation. Leur nombre et leur raffinement est encore insuffisant pour découvrir des relations complexes. De plus, de nombreuses règles redondantes sont requises afin de gérer le problème des nominalisations. Leur résolution devrait être prise en charge lors de la génération des *structures prédicat-arguments*.

Perspectives

Nous regrettons de n'avoir pu utiliser à grande échelle notre système et notamment dans le cadre de la structuration de réseaux de régulation de l'expression de gènes de cytokines. Par manque de temps, nous nous sommes restreints au développement des

méthodes et à leur utilisation sur des jeux de données qui avaient été préalablement annotées.

Les possibilités d'extension de ce travail sont nombreuses. Mis à part la résolution des limites décrites ci-dessus et l'extension de notre modèle de la régulation de gènes à d'autres données, de nouvelles directions de ce travail sont envisageables :

- Nous avons centré cette approche sur des méthodes à base de règles et d'heuristiques expertisées. Nous ne disposions pas de corpus adaptés à notre problématique lorsque ce travail a été entamé et ils ont du être constitués au fur et à mesure. Il serait désormais intéressant d'utiliser les caractéristiques des méthodes mises en œuvre comme base pour des approches d'apprentissage automatique, notamment dans le cadre de la résolution des ambigüités liées à l'IEN et pour l'acquisition des règles de conceptualisation.
- D'autre part, les relations entre entités biologiques que nous pouvons extraire sont parfois fragmentaires car nous nous positionnons à l'échelle de la phrase. En l'état actuel, les données qui ne décrivent qu'une partie des concepts que nous avons sélectionné ne sont pas utilisables. Elles ne sont cependant pas dénuées d'information mais doivent être confrontées, complétées et mises en perspective à l'échelle d'une puis de plusieurs publications scientifiques.

Disponibilité du logiciel

Le logiciel correspondant au système développé est mis à disposition du public sous la forme d'une application **Java**[5] (avec sa documentation technique) à partir de cette adresse :

`http://www.polytech.univ-nantes.fr/RelMiner/`

[5] http ://java.sun.com/

Annexe A - Penn TreeBank Project

TAB. 6 – *Parts of speech* des mots utilisés par le Penn TreeBank Project [San90]

Part of speech	Description	Exemples
“	opening quotation mark	‘ “
”	closing quotation mark	’ ”
(	opening parenthesis	([{
)	closing parenthesis	)] }
,	comma	,
-	dash	-
.	sentence terminator	.
:	colon or ellipsis	: ; ...
CC	conjunction, coordinating	& ń and both but either
CD	numeral, cardinal	mid-1890 nine-thirtyten
DT	determiner	all an another any both
EX	existential there	there
FW	foreign word	gemeinschaft jeux
IN	preposition or conjunction, subordinating	among upon whether
JJ	adjective or numeral, ordinal	third ill-mannered
JJR	adjective, comparative	bleaker busier calmer
JJS	adjective, superlative	calmest cheapest darkes
LS	list item marker	A A. First G One
MD	modal auxiliary	can couldn’t dare may
NN	noun, common, singular or mass	common-carrier cabbage
NNP	noun, proper, singular	Motown Venneboerger
NNPS	noun, proper, plural	Americans Americas
NNS	noun, common, plural	scotches bric-a-brac
PDT	pre-determiner	all both half many quite
POS	genitive marker	’ ’s
PRP	pronoun, personal	hers hisself it one
PRP$	pronoun, possessive	her his their
RB	adverb	occasionally unabatingly
RBR	adverb, comparative	further gloomier grander
RBS	adverb, superlative	best biggest second
RP	particle	aboard about go i.e.
SYM	symbol	% & ’ ” ”.)). * +
TO	”to” as preposition or infinitive marker	to
UH	interjection	Goodbye Goody Gosh
VB	verb, base form	ask assemble assess
VBD	verb, past tense	dipped pleaded swiped
VBG	verb, present participle or gerund	telegraphing stirring
VBN	verb, past participle	multihulled dilapidated
VBP	verb, present tense, not 3rd person singular	predominate wrap resort
VBZ	verb, present tense, 3rd person singular	bases reconstructs
WDT	WH-determiner	that what whatever
WP	WH-pronoun	that what whatsoever
WP$	WH-pronoun, possessive	whose
WRB	Wh-adverb	how however whence

Tab. 7 – *Parts of speech* des propositions ou des syntagmes utilisés par le Penn TreeBank Project [San90]

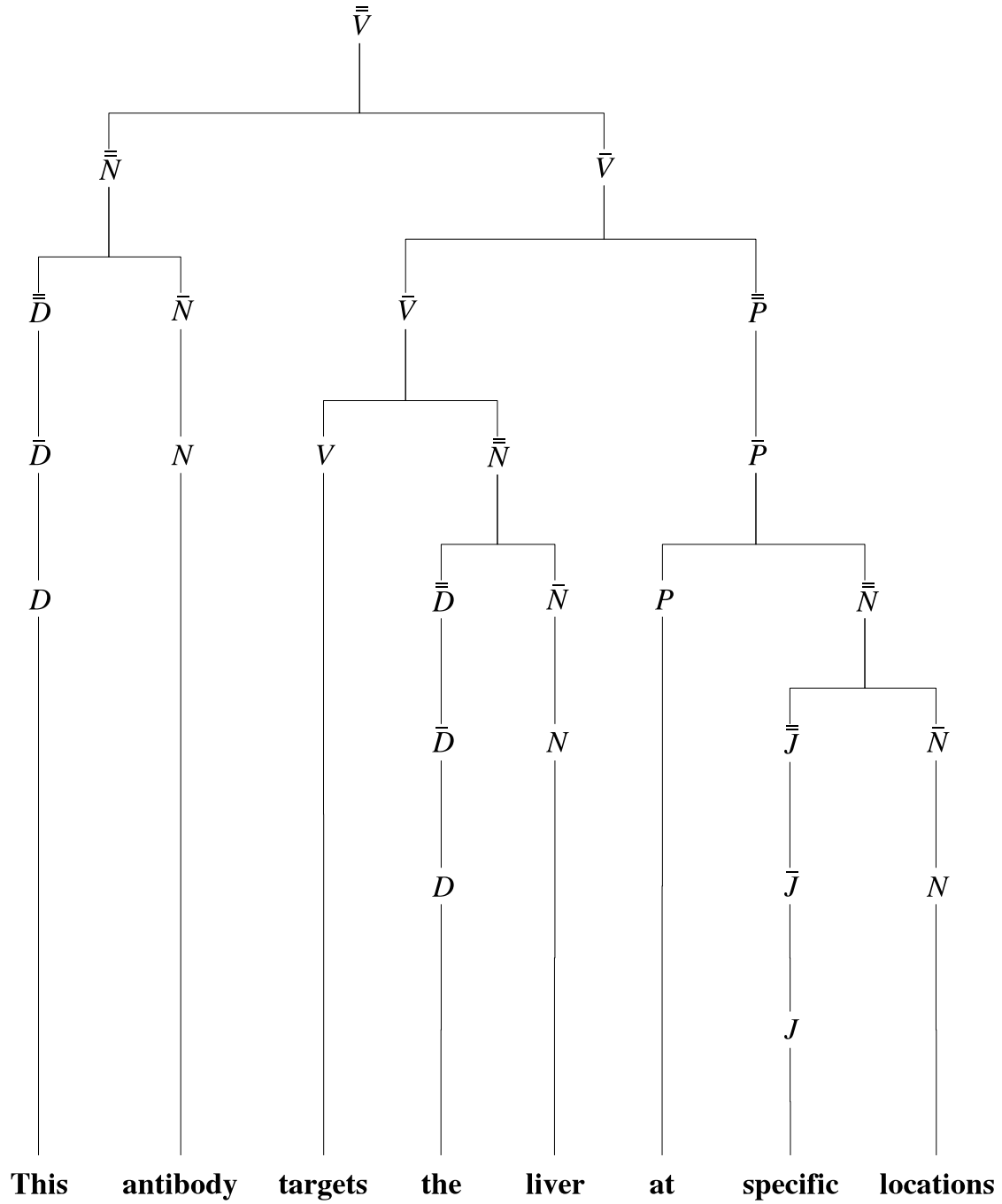
Part of speech	Description
S	Simple declarative clause
SBAR	Clause introduced by a (possibly empty) subordinating conjunction
ADJP	Adjective Phrase
ADVP	Adverb Phrase
NP	Noun Phrase
PP	Prepositional Phrase
VP	Verb Phrase
WHADJP	Wh-adjective Phrase. Adjectival phrase containing a wh-adverb, as in "how hot"
WHAVP	Wh-adverb Phrase. Adverbial phrase containing a wh-adverb such as "how" or "why"

Annexe B - Théorie de la linguistique moderne

Cette théorie prend appui sur l'hypothèse que les structures du discours de l'ensemble des langues partagent des similarités certaines. Au cœur de cette théorie, nous retrouvons plus particulièrement le schéma *X barre* [Cho70] proposé par Chomsky. Ce schéma permet de décomposer les phrases en sous-ensembles que sont les syntagmes et les mots. Il permet aussi de définir un ensemble cohérent et unifié de contraintes grammaticales.
Soient les trois méta-règles suivantes dans la version originale du schéma :
$\bar{\bar{X}} \rightarrow (Spec_{\bar{X}}), \bar{X}$
$(\bar{X} \rightarrow \bar{X}, Adjc_{\bar{X}})$
$\bar{X} \rightarrow (Comp_X ...), X$
où $X : \{P, V, N, A, J, ...\}$, P est une préposition, V un verbe, N un nom, A un adverbe, J un adjectif, $Spec$ un *spécificateur*, $Adjc$ un *ajout* et $Comp$ un *complément*. X des mots et $\bar{\bar{X}}$ et $\bar{X}$ sont des syntagmes de complexité différentes. Seul le syntagme le plus complexe spécifique d'un X est noté $\bar{\bar{X}}$, tous les autres syntagmes dépendants du même X sont notés $\bar{X}$, sans distinction. Les *spécificateurs*, *ajouts* et *compléments* sont des types particuliers de constituants, en nombre limité et spécifiques aux X ou $\bar{X}$ qu'ils décrivent. Par exemple, un déterminant est un $Spec_{\bar{N}}$, un auxiliaire est un $Spec_{\bar{V}}$ et $\bar{\bar{N}}$ est un $Comp_{\bar{X}}$. Il est a noter que la présence d'un *spécificateur* ou d'un *complément* est facultative. De la même manière, il est intéressant de voir que la troisième règle peut contenir un nombre quelconque de *compléments* et que la deuxième règle est optionnelle. Ces règles peuvent alors être combinées et représentées graphiquement sous la forme d'arbres afin de décomposer une phrase ou une de ses structures en structures plus petites.
A titre d'exemple, la figure 11 expose une des structures syntaxiques du syntagme verbal "This antibody targets the liver at specific locations", compatible avec la théorie *X barre*. La *tête* du syntagme verbal $\bar{\bar{V}}$ est le verbe "targets" qui forme un $\bar{V}$ avec le *complément*

"the liver". Une *tête* est définie comme étant le mot d'un constituant grammatical qui possède la même fonction grammaticale que l'ensemble du constituant. Le $\bar{V}$ "targets the liver" forme alors un autre $\bar{V}$ avec l'*ajout* "at specific locations". Cette théorie est extrêmement répandue et utilisée par la plupart des acteurs du TALN. Néanmoins, son unique but est de proposer une théorie globale sur la façon dont les structures d'un discours sont organisées. D'aucune manière elle ne peut répondre au problème de la compréhension du contenu, de la signification d'un discours.

FIG. 11 – Exemple d'un schéma *X barre*

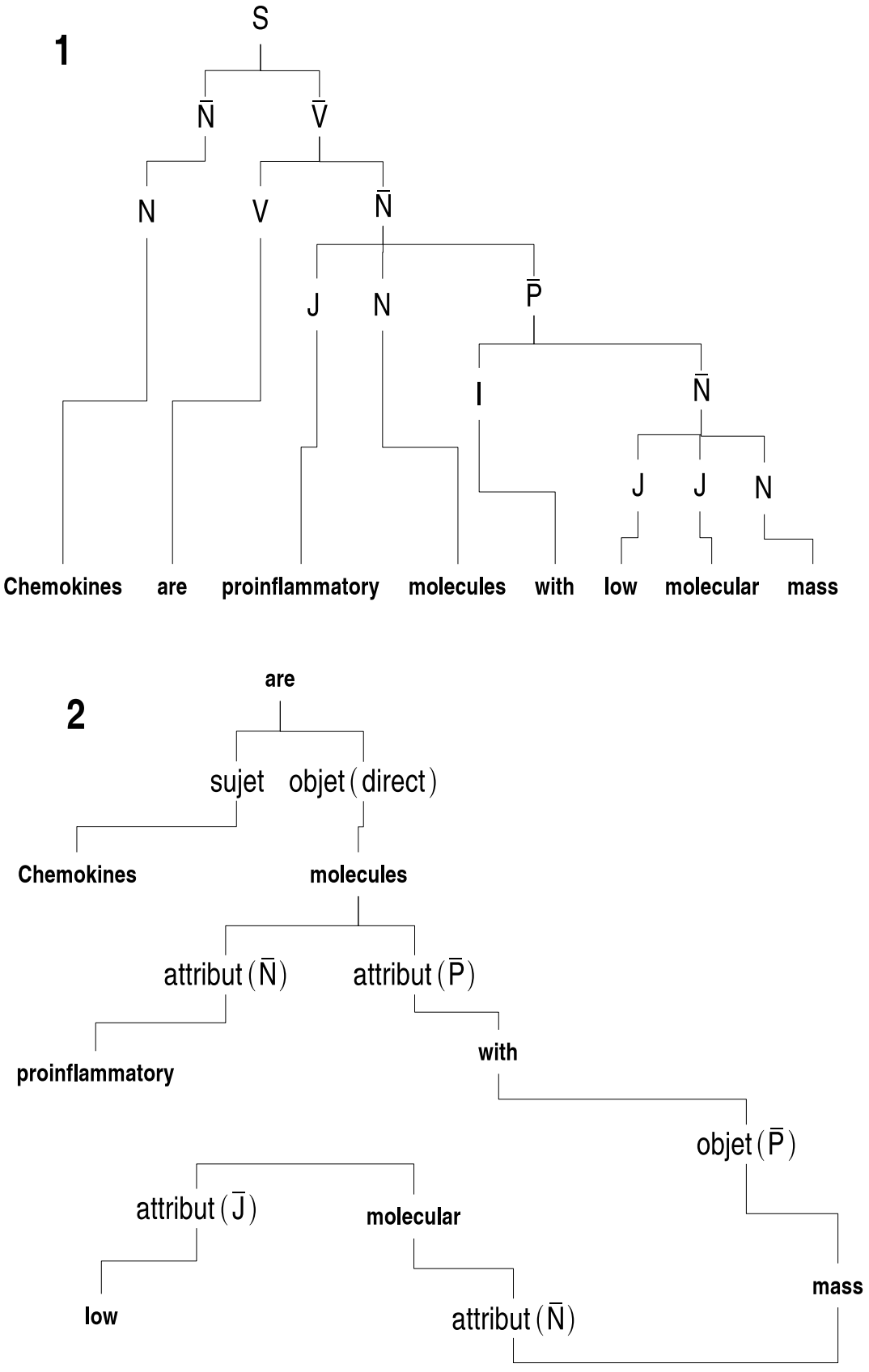


Le syntagme verbal schématisé pour l'exemple est "This antibody targets the liver at specific locations". Ceci n'est la représentation que d'une des nombreuses analyses compatibles avec la théorie *X barre*

Annexe C - Les grammaires

Une grammaire est une théorie formalisée sur les phrases et leurs structures qui sont permises (comprendre intelligibles) dans un langage donné. Une grande variété de grammaires existent, les plus connues étant certainement : les *grammaires hors contexte*, le *Tree-Adjoining Grammar (Grammaire d'arbres adjoints)*, la *Head-Driven Phrase Structure Grammar (Grammaire syntagmatique dirigée par les têtes)*, le *Generalised Phrase Structure Grammar (Grammaire syntagmatique généralisée)* et les *grammaires de dépendance*. Tous les exemples donnés, à l'exception des *grammaires de dépendance*, sont des "grammaires d'unification". C'est à dire cherchant à décomposer de manière récursive la phrase en fragments significatifs. Dans ces grammaires, la syntaxe est décrite selon des catégories prédéfinies (par exemple **S** pour une phrase et **V** pour un verbe) qui établissent des relations de composition entre elles. Néanmoins, les grammaires d'unification démontrent certaines limitations factuelles lors du processus classique de TALN. Les textes anglais contenant des fautes de grammaire ou les phrases utilisant l'adverbe 'respectively' ne peuvent être correctement représentés par ce type de grammaire. Le cas de l'utilisation de l'adverbe 'respectively' peut être étendu à toutes les constructions de phrase dîtes "distributives". Un exemple de phrase difficilement analysable par une grammaire d'unification est "IL1F5 and IL1F9 are placenta and esophagus-produced interleukins, respectively". Chomsky a défini une hiérarchie des grammaires d'unification [CS63]. Elle est formée de quatre niveaux, du plus restrictif (niveau trois) au plus large (niveau zéro). Les *grammaires hors contexte* sont présentes au niveau deux et trois. Le niveau un est celui des grammaires dépendantes du contexte. Les difficultés exposées un peu plus haut ne peuvent être résolues par le grammaires de niveau deux et trois. Néanmoins, la complexité extrême des grammaires de plus haut niveau (un et zéro) fait qu'elles ne sont pas utilisables dans des applications de TALN.

FIG. 12 – Exemples des descriptions d'une même phrase avec une *grammaire hors contexte* [1] et avec une *grammaire de dépendance* [2]



Grammaires hors contexte

Les *grammaires hors contexte* [BF69] sont utilisées à la fois par les linguistes et les informaticiens. La syntaxe de la très grande majorité des langages de programmation suit une grammaire formelle qui peut être formalisée grâce à une *grammaire hors contexte*. Pourtant, nombre de linguistes considèrent que ces grammaires sont trop pauvres pour représenter des connaissances linguistiques. Néanmoins, elles s'avèrent suffisantes pour représenter des connaissances linguistiques basiques (par exemple, celles enseignées aux jeunes enfants dans les écoles). Ces grammaires ont notamment été utilisées avec succès dans certains domaines d'application particuliers. Une *grammaire hors contexte* est une grammaire générative et transformationelle, comme définie par Chomsky, et utilise un ensemble de règles de *transformation* afin de générer une chaîne de caractères dans un langage. Ainsi, on commence par appliquer la règle de *transformation* définie par une chaîne de caractère constituée par un symbole "début". On ré-applique alors successivement les règles valides (dans un ordre non fixé, autant de fois que nécessaire) pour ré-écrire cette chaîne de caractères, jusqu'à ce que l'on ne puisse plus substituer les caractères de la chaîne. Le langage est alors l'ensemble des chaînes de caractères qui peuvent être générées de cette manière. Chaque séquence unique d'application des règles forme une chaîne particulière du langage. Par exemple, voici une grammaire qui permet de définir la syntaxe d'un langage arithmétique à trois variables très simple (en pseudo BNF) :

$$S \rightarrow a|b|c|S+S|S-S|S*S|S/S|(S)$$

Les variables étant a, b et c, S le symbole de "début" et | représentant un "ou" exclusif qui ne fait pas partie de la chaîne de caractères. Si l'on commence par appliquer la règle :
$S \rightarrow S+S$
nous remplaçons S par :
$S+S$
Si nous appliquons ensuite la règle :
$S \rightarrow (S)$
sur le premier symbole "début", puis la règle :
$S \rightarrow S*S$
sur le premier symbole "début" nous obtenons une chaîne de caractère du type :
$(S*S)+S$
Nous pouvons alors remplacer chacun des symboles "début" par des variables grâce aux

règles :
$S \rightarrow a$
$S \rightarrow b$
$S \rightarrow c$
pour donner, par exemple, la chaîne de caractères syntaxiquement correcte :
$(a * b) + c$.

Une grammaire est dite *sans contexte* lorsque l'on peut toujours appliquer une règle de *transformation* valide, et ce, quelque soit le contexte du discours. La grammaire exemple donnée plus haut est une *grammaire hors contexte*.

L'utilisation de *grammaires stochastiques sans contexte* (ici une probabilité est associée à chaque règle de *transformation*) ont montré des résultats encourageants en biologie, notamment dans la modélisation des structures secondaires de brins d'ARN [SMR+94, RE01]. En TALN, et malgré la complexité de la langue anglaise, les *grammaires hors contexte* sont utilisées en EI dans des tâches très ciblées depuis leur introduction.

Grammaires de dépendance

Les *grammaires de dépendance* [Gai65] offrent quant à elles une meilleure description des connaissances grammaticales d'un langage [Tes59]. Les mots sont reliés entre eux par couples (une relation n'est établie qu'entre deux mots). Un mot (*tête*) est relié par un lien de dépendance à d'autres mots (*modificateurs*) si, et seulement si, les contraintes d'établissement d'un lien par la *tête* sont respectées. Ces contraintes prennent alors la forme d'un ensemble de règles. Staab [Sta99] et Sleator *et al.* [ST93] décrivent cinq avantages des *grammaires de dépendance* sur les *grammaires hors contexte* :

- Une syntaxe plus simple. Seules les entités syntaxiques présentes dans le texte sont utilisées lors de l'analyse.
- Des liens sémantiques. Les liens syntaxiques sont ici très proches de liens sémantiques.
- Une lexicalisation. Les concepts utilisés sont ici 'nommés' et définis au sein d'un dictionnaire. L'adaptation de la grammaire à des spécificités d'écriture telles que retrouvées par exemple dans les documents techniques est simplifiée. La modification de la définition d'un mot au sein du dictionnaire n'affecte que la *grammaticalité* de la phrase qui le contient. Les liens établis entre les mots ne sont pas fixés par la fonction

grammaticale de ces derniers mais par leur définition au sein des dictionnaires. De plus, un mot du dictionnaire peut correspondre à plusieurs définitions au besoin.
- La discontinuité. Les relations à longue distance dans la phrase sont plus facilement capturées.
- Les *grammaires de dépendance* ne proposent aucune notion explicite de regroupement des mots dans la phrase en constituants ou catégories. Il est néanmoins aisé de retrouver la structure des constituants de la phrase à partir de la sémantique des liens. Par exemple, et grâce à un dictionnaire de la grammaire adapté à l'anglais standard, le lien de type `sujet` attache toujours un groupe nominal à un verbe.

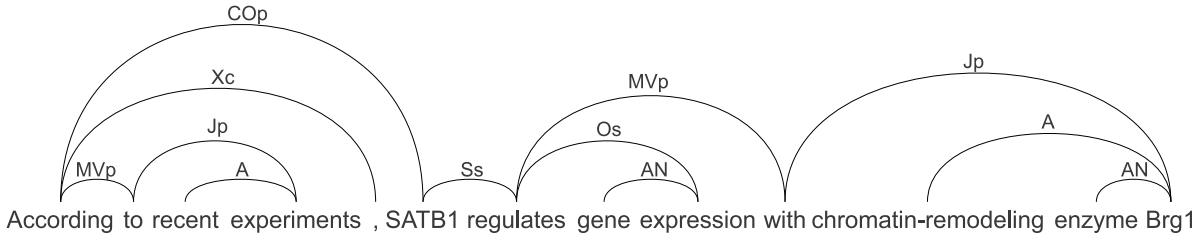
Pour toutes les raisons mentionnées, une *grammaire de dépendance* peut aussi être considérée comme une théorie sur la sémantique. L'utilisation d'une *grammaire de dépendance* nécessite obligatoirement la création préalable d'un thésaurus qui contient à la fois les contraintes d'établissement des liens spécifiques aux mots ainsi que des caractéristiques spéciales partagées par les mêmes racines de mot (voir la section 2.1.2).

Un exemple comparatif d'analyse syntaxique entre une *grammaire de dépendance* et une *grammaire hors contexte* est donné en figure 12.

Grammaire de liens

La *grammaire de lien* [ST93] est un type particulier de *grammaire de dépendance*. Il rajoute une contrainte qui n'existe pas dans une *grammaire de dépendance* classique, à savoir que les relations entre mots ont forcément une direction. Par exemple, et en langue anglaise, un verbe dans une voie active ne peut être relié par un *lien* `sujet` qu'à un nom ou pronom si ce dernier est situé à gauche du premier. Il est aussi à noter, que la recherche de validation des contraintes n'est plus imposée qu'à partir de la *tête* mais aussi à partir des *modificateurs*.

FIG. 13 – Exemple d'une analyse syntaxique réalisée en *grammaire de lien*



Une implémentation célèbre d'analyseur syntaxique compatible avec la *grammaire de*

liens porte le nom de **Link Grammar Parser (LGP)** [GLS95].

La figure 13 montre un exemple d'analyse syntaxique compatible avec la *grammaire de lien* de la phrase "According to recent experiments, SATB 1 regulates gene expression with chromatin-remodeling enzyme Brg1" et réalisée par **LGP**. Les arcs étiquetés qui connectent les mots deux-à-deux sont appelés des *liens*. Dans cette figure le mot "experiments" est lié sur sa gauche à l'adjectif "recent" avec un *lien* nommé `A`, caractéristique de la présence d'un adjectif qualificatif. Une liste de *liens* générés par **LGP** est disponible en ligne[6]. Un mot peut être lié à un autre et unique mot ou plusieurs. Par exemple, le mot "regulates" possède un *lien* `S`, qui permet de relier un nom sujet à son verbe, connecté à sa droite ainsi qu'un *lien* `O`, qui permet de relier un nom objet à son verbe, sur sa gauche. Une *grammaire de liens* est défini grâce à un dictionnaire qui contient le vocabulaire et les définitions des mots du vocabulaire. Ces définitions décrivent comment les mots peuvent être utilisés pour former des *liens*. Un *lien* est établi entre deux mots si leurs définitions sont mutuellement compatibles. Cependant, un *lien* est uniquement jugé valide s'il respecte quatre contraintes :

- le graphe des *liens* doit être plan, c'est à dire que lorsque les *liens* sont dessinés au dessus de la phrase, deux *liens* ne peuvent se croiser (contrainte de planarité),
- deux *liens* ne peuvent connecter la même paire de mots (contrainte d'exclusion),
- un mot ne peut établir plus d'une liaison par la droite avec un mot situé à sa droite, de la même manière il ne peut non plus établir plus d'un *lien* sur sa gauche avec un mot situé à sa gauche (contrainte d'ordre),
- l'ensemble des mots de la phrase doit être inter-connecté (contrainte de connectivité).

L'algorithme de **LGP** introduit la notion de *lien vide (null link)* qui permet de connecter n'importe quelle paire de mots adjacents, et ce quelque soit leur définition au sein du dictionnaire. Afin de retrouver la syntaxe sous-jacente à la phrase, dans un premier temps les *liens* entre mots sont retrouvés en suivant les contraintes grammaticales imposées par la *grammaire de liens* puis, dans un deuxième temps, les ensembles de *liens* possédant un *coût* minimal sont conservés.

Lors de l'analyse syntaxique, sur l'ensemble des combinaisons de *liens* valides possibles, seul un sous-ensemble de combinaisons est structurellement "plausible" ou "vraisemblable" en anglais standard.

[6] http ://www.link.cs.cmu.edu/link/dict/summarize-links.html

Lors de la première passe, **LGP** décompte le nombre de *liens* valides d'un graphe. Les graphes contenant le plus de *liens* sont alors considérés comme les meilleurs représentants de la structure de la phrase et grammaticalement les plus justes. Néanmoins, il n'est pas rare que la phrase ne soit pas jugée comme étant grammaticalement correcte sur toute sa longueur. Dans ce cas, l'analyse échoue car les contraintes imposées pour créer les *liens* ne sont pas respectées. Ceci est particulièrement vérifié lors de l'analyse de tran*script*ions de discussions orales, tâche pour laquelle **LGP** a été originellement conçu. Toutefois, certains 'ilots' grammaticaux peuvent être généralement repérés au sein de la phrase. Ces fragments structurels sont d'une importance cruciale car ils permettent, à défaut d'obtenir une structure grammaticale globale de la phrase, de comprendre de façon parcellaire la syntaxe et de ne pas perdre l'intégralité de l'information contenue dans la phrase. **LGP** utilise un algorithme d'analyse syntaxique robuste qui lui permet si la première passe aboutit à une impasse grammaticale, de réaliser une deuxième passe à l'aide des *liens* virtuels que sont les *liens vides*.

La notion de *coût* d'un graphe est alors introduite par **LGP** afin de déterminer les différents sous-ensembles de graphes de *liens* valides d'une phrase. La mesure de *coût* définie dans **LGP** est une heuristique qui représente, de manière informelle, la *grammaticalité*. Le *coût* est une mesure absolue qui est associée à chaque ensemble de *liens* qui respecte les contraintes de planarité et d'exclusion. Un *coût* plus faible reflète une meilleure *grammaticalité*. Comme toute heuristique, chaque fonction de *coût* proposée peut générer des exemples de phrases structurellement meilleures avec des *coûts* plus élevés. Les auteurs ont mis au point la fonction de *coût* et réalisés des expérimentations de robustesse sur le corpus **Switchboard**[7]. Ce corpus contient la tran*script*ion écrite de plus de 150h de conversation téléphonique portant sur 70 thèmes différents.

FIG. 14 – Exemple d'un graphe de *liens* obtenu par **LGP** et contenant deux *liens vides*

Le *coût* d'un graphe dépend du nombre de *liens vides* qui le composent et favorise les graphes contenant le plus de *liens* par rapport aux graphes plus pauvres. La figure 14 montre un exemple de graphe de *liens* contenant deux *liens vides* dessinés sous la forme d'arcs hachurés. Le *coût* d'un graphe est défini comme étant égal au nombre de *liens vides*

[7] http ://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp ?catalogId=LDC93T4

qui le compose. la valeur obtenue est ainsi bornée entre les nombre de mots de la phrase - 1 (graphe sans *lien*) et 0 (graphe complet). Les graphes qui sont sensés représenter au mieux la syntaxe sous-jacente sont ainsi les graphes associés aux *coûts* les plus faibles.

Bibliographie

[A.99] Travers A. An engine for nucleosome remodeling. *Cell*, 3(96) :311–314, 1999.

[Abn96] S. Abney. Partial parsing via finite state cascades. *Natural Language Engineering*, 2(4) :337–344, 1996.

[ACH+00] A. Ar, B. Chang, H. Humphrey, S. Mork, J. Nelson, S. Rindflesch, and T. Wilbur. The nlm indexing initiative. In *Proc AMIA Symp*, pages 17–21, 2000.

[AD00] Wolffe A. and Guschin D. Review : chromatin structural features and targets that regulate transcription. *J Struct Biol*, 2-3(129) :102–122, 2000.

[Alo07] U. Alon. Network motifs : theory and experimental approaches. *Nature Reviews Genetics*, 8(6) :450–461, 2007.

[ALP+00] T. Agalioti, S. Lomvardas, B. Parekh, J. Yie, T. Maniatis, and D. Thanos. Ordered recruitment of chromatin modifying and general transcription factors to the ifn-beta promoter. *Cell*, 103(4) :667–678, 2000.

[AM92] H. Alshawi and R. Moore. Introduction to the cle. In H. Alshawi, editor, *The Core Language Engine*, pages 1–10. MIT Press, 1992.

[AMS96] A. Abbas, K. Murphy, and A. Sher. Functional diversity of helper t lymphocytes. *Nature*, 383(6603) :787–93, 1996.

[AR98] S. Agarwal and A. Rao. Modulation of chromatin structure regulates cytokine gene expression during t cell differentiation. *Immunity*, 9(6) :765–75, 1998.

[Aro01] A. R. Aronson. Effective mapping of biomedical text to the umls metathesaurus : the metamap program. In *Proc AMIA Symp.*, pages 17–21, 2001.

[BA00] A. Bairoch and R. Apweiler. The SWISS-PROT protein sequence database and its supplement trEMBL in 2000. *Nucleic Acids Res.*, 28 :45–48, 2000.

[BACV99] C. Blaschke, M. Andrade, C., and A. Valencia. Automatic extraction of biological information from scientific text : Protein-protein interactions. In *Proc Int Conf Intell Syst Mol Biol*, pages 60–67, 1999.

[Bal97] B. Baldwin. Cogniac : High precision coreference with limited knowlege and linguistic resources. In *Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts (ACL-97 workshop)*, pages 38–45, 1997.

[BAOV99] C. Blaschke, M. Andrade, C. Ouzounis, and A. Valencia. Automatic extraction of biological information from scientific text : protein-protein interactions. In *Proc Int Conf Intell Syst Mol Biol*, pages 60–67, 1999.

[BBH03] G. Bader, D. Betel, and C. Hogue. Bind : the biomolecular interaction network database. *Nucleic Acids Res*, 31(1) :248–250, January 2003.

[BF69] D. Bobrow and J. Fraser. An augmented state transition network analysis procedure. In *IJCAI*, pages 557–568, 1969.

[BH97] M. Brown and J. Hural. Functions of il-4 and control of its expression. *Crit Rev Immunol*, 17(1) :1–32, 1997.

[BM05] R. Bunescu and R. Mooney. A shortest path dependency kernel for relation extraction. In *Proceedings of the Joint Conference on Human Language Technology / Empirical Methods in Natural Language Processing (HLT/EMNLP)*, pages 724–731, October 2005.

[BMRJ+05] K. Brettingham-Moore, S. Rao, T. Juelich, M. Shannon, and A. Holloway. Gm-csf promoter chromatin remodelling and gene transcription display distinct signal and transcription factor requirements. *Nucleic Acids Res*, 33(1) :225–34, 2005.

[BRR+05] R. Bunescu, G. Ruifang, K. Rohit, E. Marcotte, R. Mooney, A. Ramani, and Y. Wong. Comparative experiments on learning information extractors for proteins and their interactions. *Artificial Intelligence in Medicine*, 33(2) :139–155, February 2005.

[BSAG98] A. Borthwick, J. Sterling, E. Agichtein, and R. Grishman. Exploiting diverse knowledge sources via maximum entropy in named entity recognition. In *Sixth Workshop on Very Large Corpora*. Association for Computational Linguistics, 1998.

[BV02] C. Blaschke and A. Valencia. The frame-based module of the suiseki information extraction system. *IEEE Intelligent Systems*, 17(2) :14–20, 2002.

[Car98] M. Carey. The enhanceosome and transcriptional synergy. *Cell*, 92(1) :5–8, January 1998.

[CDF+04] R. Christopher, A. Dhiman, J. Fox, R. Gendelman, T. Haberitcher, D. Kagle, G. Spizz, I. Khalil, and C. Hill. Data-driven computer simulation of human cancer cell. *Ann N Y Acad Sci*, 1020, 2004.

[Cho70] N. Chomsky. Remarks on nominalization. In *Readings in Transformational Grammar*, pages 184–221. Ginn and co., 1970.

[CK99] M. Craven and J. Kumlien. Constructing biological knowledge bases by extracting information from text sources. In *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*, pages 77–86. AAAI Press, 1999.

[CN07] A. Casamassimi and C. Napoli. Mediator complexes and eukaryotic transcription regulation : An overview. *Biochimie*, August 2007.

[CNT00] N. Collier, C. No, and J. Tsujii. Extracting the names of genes and gene products with a hidden markov model. In *Proc. COLING 2000*, pages 201–207, 2000.

[CS63] N. Chomsky and M. Schützenberger. The algebraic theory of context-free languages. In P. Braffort and D. Hirschberg, editors, *Computer Programming and Formal Systems*, pages 118–161. noho, 1963.

[CSB+93] P. Cockerill, M. Shannon, A. Bert, G. Ryan, and M. Vadas. The granulocyte-macrophage colony-stimulating factor/interleukin 3 locus is regulated by an inducible cyclosporin a-sensitive enhancer. In *Proc Natl Acad Sci USA*, volume 90, page 2466Ű2470, March 1993.

[D.95] Yarowsky D. Unsupervised word sense disambiguation rivaling supervised methods. In *Meeting of the Association for Computational Linguistics*, pages 189–196, 1995.

[DC05] R. Drysdale and M. Crosby. Flybase : genes and gene models. *Nucleic Acids Research*, 33(Supplement 1) :D390+, January 2005.

[DCL+94] S. Dunn, L. Coles, R. Lang, S. Gerondakis, M. Vadas, and M. Shannon. Requirement for nuclear factor (nf)-kappa b p65 and nf-interleukin-6 binding elements in the tumor necrosis factor response region of the granulocyte colony-stimulating factor promoter. *Blood*, 83(9) :2469–2479, 1994.

[DOTW97] S. Davidson, G. Overton, V. Tannen, and L. Wong. Biokleisli : A digital library for biomedical researchers. *Int. J. on Digital Libraries*, 1(1) :36–53, 1997.

[DYE+04] N. Daraselia, A. Yuryev, S. Egorov, S. Novichkova, A. Nikitin, and I. Mazo. Extracting human protein interactions from medline using a full-sentence parser. *Bioinformatics*, 20(5) :604–611, March 2004.

[EBK+05] J. Eppig, C. Bult, J. Kadin, J. Richardson, and J. Blake. The mouse genome database (mgd) : from genes to mice-a community resource for mouse biology. *Nucleic Acids Research*, 33(Supplement 1) :D471+, January 2005.

[EOR07] G. Erkan, A. Ozgur, and D. Radev. Semi-supervised classification for extracting protein interaction sentences using dependency parsing. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 228–237, 2007.

[FAM00] K. Frantzi, S. Ananiadou, and H. Mima. Automatic recognition of multi-word terms :. the c-value/nc-value method. *International Journal on Digital Libraries*, V3(2) :115–130, 2000.

[Fil68] C. Fillmore. The case for case. In E. Bach and R. Harms, editors, *Universals in Linguistic Theory*, pages 1–88. Holt, Rinehart and Winston, Inc., 1968.

[FTTT98] K. Fukuda, T. Tsunoda, A. Tamura, and T. Takagi. Toward information extraction : Identifying protein names from biological papers. In *Proc. of the Pacific Symposium on Biocomputing '98*, 1998.

[FUB+00] J. Falvo, A. Uglialoro, B. Brinkman, M. Merika, B. Parekh, E. Tsai, H. King, A. Morielli, E. Peralta, T. Maniatis, D. Thanos, and A. Goldfeld. Stimulus-specific assembly of enhancer complexes on the tumor necrosis factor alpha gene promoter. *Mol Cell Biol*, 20(6) :2239–2247, 2000.

[Gai65] H. Gaifman. Dependency systems and phrase-structure systems. *Information and Control*, 8(3) :304–337, 1965.

[GL07] J. Goutsias and N. Lee. Computational and experimental approaches for modeling gene regulatory networks. *Curr Pharm Des*, 13(14) :1415–1436, 2007.

[GLS95] D. Grinberg, J. Lafferty, and D. Sleator. A robust parsing algorithm for LINK grammars. Technical Report CMU-CS-TR-95-125, CMU, 1995.

[HDG00] K. Humphreys, G. Demetriou, and R. Gaizauskas. Two applications of information extraction to biological science journal articles : enzyme interactions and protein structures. *Pac Symp Biocomput*, pages 505–516, 2000.

[HDR01] V. Hatzivassiloglou, P. Duboué, and A. Rzhetsky. Disambiguating proteins, genes, and rna in text : a machine learning approach. *Bioinformatics*, 17(1) :97–106, 2001.

[Hea92] M. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics*, pages 539–545, 1992.

[Her05] W. Hersh. Report on the trec 2004 genomics track. *SIGIR Forum*, 39(1) :21–24, 2005.

[HFMZ03] D. Hanisch, J. Fluck, H. Mevissen, and R. Zimmer. Playing biology's name game : Identifying protein names in scientific text. In *Pacific Symposium on Biocomputing*, pages 403–414, 2003.

[HMH+04] Z. Hu, I. Mani, V. Hermoso, H. Liu, and C. Wu. iprolink : an integrated protein resource for literature mining. *Comput Biol Chem*, 28(5-6) :409–416, December 2004.

[IEO01] I. Iliopoulos, A. Enright, and C. Ouzounis. Textquest : document clustering of medline abstracts for concept discovery in molecular biology. In *Pac Symp Biocomput*, pages 384–395, 2001.

[IS89] R. Ingria and D. Stallard. A computational mechanism for pronominal reference. In *Proceedings of the 27th annual meeting on Association for Computational Linguistics*, pages 262–271, 1989.

[Jac01] C. Jacquemin. *Spotting and Discovering Terms through NLP*. MIT Press, Cambridge MA, 2001.

[JDOTT03] K. Jin-Dong, T. Ohta, Y. Teteisi, and J. Tsujii. Genia corpus - a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl. 1) :i180–i182, 2003.

[JLC+94] Chen J., Attardi L., Verrijzer C., Yokomori K., and Tjian R. Assembly of recombinant tfiid reveals differential coactivator requirements for distinct transcriptional activators. *Cell*, 1(79) :93–105, Oct 1994.

[JLKH01] T. Jenssen, A. Laegreid, J. Komorowski, and E. Hovig. A literature network of human genes for high-throughput analysis of gene expression. *Nat Genet*, 28(1) :21–28, May 2001.

[JLR95] J. Jain, C. Loh, and A. Rao. Transcriptional regulation of the il-2 gene. *Current Opinion in Immunology*, 7(3) :333–342, 1995.

[JM03] P. Jackson and I. Moulinier. Briefly noted : natural language processing for online applications : Text retrieval, extraction, and categorization. *Computational Linguistics*, 29(3) :510–511, 2003.

[JMM+93] J. Jain, P. McCaffrey, Z. Miner, T. Kerppola, J. Lambert, G. Verdine, T. Curran, and A. Rao. The t-cell transcription factor nfatp is a substrate for calcineurin and interacts with fos and jun. *Nature*, 365(6444) :352–355, 1993.

[Kar00] P. Karp. An ontology for biological function based on molecular interactions. *Bioinformatics*, 16(3) :269–285, Mar 2000.

[Kaz02] Jun'ichi Kazama. Named entity recognition in the molecular-biology domain. In *Workshop on Natural Language Processing and Ontology Building in Biology*, 2002.

[KB96] C. Kennedy and B. Boguraev. Anaphora for everyone : Pronominal anaphora resolution without a parser. In *Proceedings of the 16th International Conference on Computational Linguistics, COLING-96*, pages 113–118, 1996.

[KCE04] D. Kightley, N. Chandra, and K. Elliston. Inferring gene regulatory networks from raw data : A molecular epistemics approach. In *Pacific Symposium on Biocomputing*, pages 510–520, 2004.

[KCMT87] J. Kadonaga, K. Carner, F. Masiarz, and R. Tjian. Isolation of cdna encoding transcription factor sp1 and functional analysis of the dna binding domain. *Cell*, 51 :1079–1090, Dec 1987.

[KGF+02] Franzén K., Eriksson G., Olsson F., Asker L., Lidén P., and Cöster J. Protein names and how to find them. *Int J Med Inform*, 67(1-3) :49–61, December 2002.

[KI05] R. Kelley and T. Ideker. Systematic interpretation of genetic interactions using protein networks. *Nature Biotechnology*, 23(5) :561–566, May 2005.

[KKMBW99] A. Kel, O. Kel-Margoulis, V. Babenko, and E. Wingender. Recognition of nfatp/ap-1 composite elements within genes induced upon the activation of immune cells. *J Mol Biol*, 288(3) :353–376, 1999.

[KKT03] A. Koike, Y. Kobayashi, and T. Takagi. Kinase pathway database : An integrated protein-kinase and nlp-based protein-interaction resource. *Genome Res.*, 13(6A) :1231–43, 2003.

[KL99] R. Kornberg and Y Lorch. Twenty-five years of the nucleosome, fundamental particle of the eukaryotic chromosome. *Cell*, (98) :285Ű294, 1999.

[KLRS94] C. Kunsch, R. Lang, C. Rosen, and M. Shannon. Synergistic transcriptional activation of the il-8 gene by nf-kappa b p65 (rela) and nf-il-6. *J Immunol*, 153(1) :153–164, 1994.

[KRMF00] M. Krauthammer, A. Rzhetsky, P. Morozov, and C. Friedman. Using blast for identifying gene and protein names in journal articles. *Gene*, 259(1-2) :245–252, December 2000.

[KSF+02] W. Kent, C. Sugnet, T. Furey, K. Roskin, T. Pringle, A. Zahler, and D. Haussler. The human genome browser at ucsc. *Genome Res*, 12(6) :996–1006, Jun 2002.

[KT04] A. Koike and T. Takagi. Gene/protein/family name recognition in biomedical literature. In *Proceedings of HLT/NAACL BioLINK workshop*, pages 9–16, 2004.

[Lap02] M. Lapata. The disambiguation of nominalisations. *Computational Linguistics*, 28(3) :357–388, 2002.

[Lar04] Pierre Larrivee. Le groupe nominal epithete : Monsieur le linguiste, mes amis les linguistes, les linguistes mes amis. *Linguisticae Investigationes*, 27 :47–81, 2004.

[LCM+03] G. Leroy, H. Chen, J. Martinez, S. Eggers, R. Falsey, K. Kislin, Z. Huang, J. Li, J. Xu, D. McDonald, and T. Ng. Genescene : Biomedical text and data mining. In *JCDL*, pages 116–118, 2003.

[LHC04] C. Lee, W. Hou, and H. Chen. Annotating multiple types of biomedical entities : A single word classificication approach. In *Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, 2004.

[LHW03] R. McEntire L. Hirschman, C. Friedman and C. Wu. Linking biological language information and knowledge. In *Pacific Symposium on Biocomputing*, pages 439–450, 2003.

[Lin90] C. Lindberg. The unified medical language system (umls) of the national library of medicine. *J Am Med Rec Assoc*, 60 :40–42, 1990.

[LL94] S. Lappin and H. Leass. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20 :535–561, 1994.

[McK98] V. McKusick. Mendelian inheritance in man. a catalog of human genes and genetic disorders. *Baltimore : Johns Hopkins University Press*, 12th edition, 1998.

[MCSM04] D. Mcdonald, H. Chen, H. Su, and B. Marshall. Extracting gene pathway relations using a hybrid grammar : the arizona relation parser. *Bioinformatics*, 20(18) :3370–3378, 2004.

[MGE01] Yui M., Hernández-Hoyos G., and Rothenberg E. A new regulatory region of the il-2 locus that confers position-independent transgene expression. *J Immunol*, 3(166) :1730–1739, 2001.

[Mic05] J. Mickolajczak. *Extraction de signatures complexes pour la découverte de nouveaux membres dans des familles de protéines connues.* PhD thesis, 2005.

[Mik99] A. Mikheev. Periods, capitalized words, etc. *Computational Linguistics*, page 25, 1999.

[Min75] M. Minsky. Minsky's frame system theory. In *TINLAP '75 : Proceedings of the 1975 workshop on Theoretical issues in natural language processing*, pages 104–116. Association for Computational Linguistics, 1975.

[Mit99] R. Mitkov. Anaphora resolution : the state of the art. In *COLING'98/ACL'98 tutorial on anaphora resolution*, 1999.

[MKMF+06] V. Matys, O. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, A. Barre-Dirrie, I. Reuter, D. Chekmenev, M. Krull, K. Hornischer, N. Voss, P. Stegmaier, B. Lewicki-Potapov, H. Saxel, A. Kel, and E. Wingender. Transfac and its module transcompel : transcriptional gene regulation in eukaryotes. *Nucleic Acids Res*, 34(Database issue) :D108–10, 2006.

[MM87] Bodner M. and Karin M. A pituitary-specific trans-acting factor can stimulate transcription from the growth hormone promoter in extracts of nonexpressing cells. *Cell*, 50(2) :267Ű275, Jul 1987.

[MSOI+02] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs : simple building blocks of complex networks. *Science*, 298(5594) :824–827, October 2002.

[MTK99] Cosma M., Tanaka T., and Nasmyth K. Ordered recruitment of transcription and chromatin remodeling factors to a cell cycle- and developmentally regulated promoter. *Cell*, 3(97) :299–311, April 1999.

[NCT00] C. Nobata, N. Collier, and J. Tsujii. Comparison between tagged corpora for the named entity task. In *Association for Computational Linguistics*, 2000.

[NM98] H. Nakagawa and T. Mori. Nested collocation and compound noun for term recognition. In *First Workshop on Computational Terminology COMPTERM'98*, pages 64–70, 1998.

[NMB03] S. Nirenburg, M. McShane, and S. Beale. Enhancing recall in information extraction through ontological semantics. In *Workshop on Ontologies and Information Extraction*, 2003.

[Nob88] H. Noble. *Natural language processing*. Blackwell Scientific Publications, 1988.

[NRVS03] M. Narayanaswamy, K. Ravikumar, and K. Vijay-Shanker. A biological named entity recognizer. In *Pac Symp Biocomput*, pages 427–438, 2003.

[NSA03] G. Nenadic, I. Spasic, and S. Ananiadou. Terminology-driven mining of biomedical literature. *Bioinformatics*, 19(8) :938–943, May 2003.

[NW99] S. Ng and M. Wong. Toward routine automatic pathway discovery from on-line scientific text abstracts. In *Genome Inform Ser Workshop Genome Inform*, volume 10, pages 104–112, 1999.

[OHTT01] T. Ono, H. Hishigaki, A. Tanigami, and T. Takagi. Automated extraction of information on protein-protein interactions from the biological literature. *Bioinformatics*, 17(2) :155–161, February 2001.

[ORS+02] D. Oliver, D. Rubin, J. Stuart, M. Hewett, T. Klein, and R. Altman. Ontology development for a pharmacogenetics knowledge base. In *Pacific Symposium on Biocomputing*, pages 65–76, 2002.

[OTC+00] T. Ohta, Y. Tateishi, N. Collier, C. Nobata, and J. Tsujii. Building an annotated corpus from biology research papaers. In *COLING-2000 Workshop on Semantic Annotation and Intelligent Content*, pages 28–34, 2000.

[PB01] Yo. Park and R. Byrd. Hybrid text mining for finding abbreviations and their definitions. In *2001 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 126–133, 2001.

[PCC+01] J. Pustejovsky, J. Castaño, B. Cochran, M. Kotecki, and M. Morrell. Automatic extraction of acronym-meaning pairs from medline databases. *Medinfo*, 10(Pt 1) :371–375, 2001.

[PCG+04] R. Podowski, J. Cleary, N. Goncharoff, G. Amoutzias, and W. Hayes. Azure, a scalable system for automated term disambiguation of gene and protein names. In *Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference (CSB'04)*, pages 415–424, 2004.

[PCP00] Cheung P., Allis C., and Sassone-Corsi P. Signaling to chromatin through histone modifications. *Cell*, 2(103) :263–271, 2000.

[PCZ+02] J. Pustejovsky, J. Castaño, J. Zhang, M. Kotecki, and B. Cochran. Robust relational parsing over biomedical literature : extracting inhibit relations. *Pac Symp Biocomput*, pages 362–373, 2002.

[PGH+07] S. Pyysalo, F. Ginter, J. Heimonen, J. Bjorne, J. Boberg, J. Jarvinen, and T. Salakoski. Bioinfer : A corpus for information extraction in the biomedical domain. *BMC Bioinformatics*, 8(1), 2007.

[PGP+04] S. Pyysalo, F. Ginter, T. Pahikkala, J. Boberg, J. Järvinen, T. Salakoski, and J. Koivula. Analysis of link grammar on biomedical dependency corpus targeted at protein-protein interactions. In *COLING 2004 International Joint workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP) 2004*, pages 15–21, August 28th and 29th 2004.

[Pin89] S. Pinker. *Learnability and Cognition : The acquisition of argument structure*. MIT Press, 1989.

[PKK01] J. Park, H. Kim, and J. Kim. Bidirectional incremental parsing for automatic pathway identication with combinatory categorial grammar. In *Pacic Rim Symposium on Biocomputing*, 2001.

[PKW+00] Cheung P., Tanner K., Cheung W., Sassone-Corsi P., Denu J., and Allis C. Synergistic coupling of histone h3 phosphorylation and acetylation in response to epidermal growth factor stimulation. *Mol Cell*, 6(5) :905–915, 2000.

[PM01] K. Pruitt and D. Maglott. Refseq and locuslink : Ncbi gene-centered resources. *Nucleic Acids Research*, 29(1) :137–140, 2001.

[Por80] M. Porter. An algorithm for suffix stripping. *Program*, 14(3) :130–137, 1980.

[PSAN06] S. Pyysalo, T. Salakoski, S. Aubin, and A. Nazarenko. Lexical adaptation of link grammar to the biomedical sublanguage : a comparative evaluation of three approaches. *BMC Bioinformatics*, 7(Suppl 3), 2006.

[PW05] M. Péry-Woodley. *Sémantique et corpus*. Condamines, A., 2005.

[PZC+04] H. Pan, L. Zuo, V. Choudhary, Z. Zhang, S. Leow, F. Chong, Y. Huang, V. Ong, B. Mohanty, S. Tan, S. Krishnan, and V. Bajic. Dragon tf association miner : a system for exploring transcription factor associations through text-mining. *Nucleic Acids Res*, 32(Web Server issue), July 2004.

[QKC04] Long Q., Min-Yen K., and Tat-Seng C. A public reference implementation of the rap anaphora resolution algorithm. *CoRR*, cs.CL/0406031, 2004.

[Ram87] A. Ram. Aqua : Asking questions and understanding answers. In *AAAI*, pages 312–316, 1987.

[RCSA02] S. Raychaudhuri, J. Chang, P. Sutphin, and R. Altman. Associating genes with gene ontology codes using a maximum entropy analysis of biomedical literature. *Genome Res*, 12(1) :203–214, January 2002.

[RE01] E. Rivas and S. Eddy. Noncoding rna gene detection using comparative sequence analysis. *BMC Bioinformatics*, 2(1), 2001.

[RHG95] J. Rooney, T. Hoey, and L. Glimcher. Coordinate and cooperative roles for nf-at and ap-1 in the regulation of the murine il-4 gene. *Immunity*, 2(5) :473–483, 1995.

[RS99] S. Reiner and R. Seder. Dealing from the evolutionary pawnshop : how lymphocytes make decisions. *Immunity*, 11(1) :1–10, 1999.

[RTWH00] T. Rindflesch, L. Tanabe, J. Weinstein, and L. Hunter. Edgar : extraction of drugs, genes and relations from the biomedical literature. *Pac Symp Biocomput*, pages 517–528, 2000.

[S.97] Gilbert S. *Developmental Biology, Fifth Edition*. Sinauer Associates, Inc., 1997.

[San90] B. Santorini. *Part-of-speech tagging guidelines for the Penn Treebank Project*. 1990. Technical report MS-CIS-90-47.

[SB00] B. Stapley and G. Benoit. Biobibliometrics : information retrieval and visualization from co-occurrences of gene names in medline abstracts. In *Pac Symp Biocomput*, pages 529–540, 2000.

[SC03] M. Skounakis and M. Craven. Evidence combination in biomedical natural-language processing. In *BIOKDD*, pages 25–32, 2003.

[SCR03] M. Skounakis, M. Craven, and S. Ray. Hierarchical hidden markov models for information extraction. In *IJCAI*, pages 427–433, 2003.

[SEW05] M. Scherf, A. Epple, and T. Werner. The next generation of literature analysis : integration of genomic analysis into text mining. *Brief Bioinform*, 6(3) :287–297, September 2005.

[SEWB00] H. Shatkay, S. Edwards, W. Wilbur, and M. Boguski. Genes, themes and microarrays : using information retrieval for large-scale gene analysis. In *Proc Int Conf Intell Syst Mol Biol*, volume 8, pages 317–328, 2000.

[SF03] H. Shatkay and R. Feldman. Mining the biomedical literature in the genomic era : an overview. *J Comput Biol*, 10(6) :821–855, 2003.

[SH03] A. Schwartz and M. Hearst. A simple algorithm for identifying abbreviation definitions in biomedical texts. In *Proceedings of the Pacific Symposium on Biocomputing*, 2003.

[SJO+06] J. Saric, L. Jensen, R. Ouzounova, I. Rojas, and P. Bork. Extraction of regulatory gene/protein networks from medline. *Bioinformatics*, 22(6) :645–650, 2006.

[SL04] P. Srinivasan and B. Libbus. Mining medline for implicit links between dietary substances and diseases. *Bioinformatics*, 20(1) :290–296, 2004.

[SMR+94] Y. Sakakibara, M., R., I. Mian, K. Sjölander, R. Underwood, and D. Haussler. Stochastic context-free grammars for tRNA modeling. *Nucleic Acids Research*, 22 :5112–5120, 1994.

[SMS+04] L. Salwinski, C. Miller, A. Smith, F. Pettit, J. Bowie, and D. Eisenberg. The database of interacting proteins : 2004 update. *Nucleic Acids Research*, 32(Database issue D449-D451), 2004.

[SPT98] T. Sekimizu, H. Park, and J. Tsujii. Identifying the interaction between genes and gene products based on frequently seen verbs in medline abstracts. In *Genome Inform Ser Workshop Genome Inform*, 1998.

[ST93] D. Sleator and D. Temperley. Parsing English with a link grammar. In *Third International Workshop on Parsing Technologies*, 1993.

[Sta99] S. Staab. *Grading Knowledge, Extracting Degree Information from Texts*, volume 1744 of *Lecture Notes in Computer Science*. Springer, 1999.

[Sun93] B. Sundheim. The message understanding conferences. In *Proceedings of a workshop on held at Fredericksburg, Virginia*, pages 5–5. Association for Computational Linguistics, 1993.

[SWS+04] M. Schuemie, M. Weeber, B. Schijvenaars, E. van Mulligen, C. van der Eijk, R. Jelier, B. Mons, and J. Kors. Distribution of information in biomedical abstracts and full-text publications. *Bioinformatics*, 20(16) :2597–2604, November 2004.

[T.07] Ideker T. Network genomics. In *Ernst Schering Res Found Workshop*, volume 61, pages 89–115, 2007.

[TCL+04] O. Tuason, L. Chen, H. Liu, J. Blake, and C. Friedman. Biological nomenclatures : Source of lexical knowledge and ambiguity. In *Proceedings of the Pacific Symposium of Biocomputing*, number 9, pages 238–249, 2004.

[Tes59] L. Tesnière. *Eléments de syntaxe structurale*. Klincksieck, 1959.

[TFT+00] E. Tsai, J. Falvo, A. Tsytsykova, A. Barczak, A. Reimold, L. Glimcher, M. Fenton, D. Gordon, I. Dunn, and A. Goldfeld. A lipopolysaccharide-specific enhancer complex involving ets, elk-1, sp1, and creb binding protein and p300 is recruited to the tumor necrosis factor alpha promoter in vivo. *Mol Cell Biol*, 20(16) :6084–6094, 2000.

[TG03] J. Temkin and M. Gilder. Extraction of protein interaction information from unstructured text using a context-free grammar. *Bioinformatics*, 19(16) :2046–2053, November 2003.

[TJP+96] E. Tsai, J. Jain, P. Pesavento, A. Rao, and A. Goldfeld. Tumor necrosis factor alpha gene regulation in activated t cells involves atf-2/jun and nfatp. *Mol Cell Biol*, 16(2) :459–467, 1996.

[TKMS03] K. Toutanova, D. Klein, C. Manning, and Y. Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *HLT-NAACL*, 2003.

[TMOP00] J. Thomas, D. Milward, C. Ouzounis, and S. Pulman. Automatic extraction of protein interactions from scientific abstracts. In *Pac Symp Biocomput*, pages 541–552, 2000.

[TSS+99] L. Tanabe, U. Scherf, L. Smith, J. Lee, L. Hunter, and J. Weinstein. Medminer : an internet text-mining tool for biomedical information, with application to gene expression profiling. *Biotechniques*, 27(6), December 1999.

[TT04] Y. Tsuruoka and J. Tsujii. Boosting precision and recall of dictionary-based protein name recognition. In *Proceedings of the ACL-03 Workshop on Natural Language Processing in Biomedicine*, 2004.

[TTJD+05] Y. Tsuruoka, Y. Tateishi, K. Jin-Dong, T. Ohta, J. McNaught, S. Ananiadou, and J. Tsujii. Developing a robust part-of-speech tagger for biomedical text. In *Proceedings of the 10th Panhellenic Conference on Informatics*, 2005.

[TTS+95] R. Thomas, M. Tymms, A. Seth, M. Shannon, and I. Kola. Ets1 transactivates the human gm-csf promoter in jurkat t cells stimulated with pma and ionomycin. *Oncogene*, 11(10) :2135–2443, 1995.

[TV06] J. Tamames and A. Valencia. The success (or not) of hugo nomenclature. *Genome Biology*, 7(5) :402, 2006.

[TYTG96] E. Tsai, J. Yie, D. Thanos, and A. Goldfeld. Cell-type-specific regulation of the human tumor necrosis factor alpha gene in b cells and t cells by nfatp and atf-2/jun. *Mol Cell Biol*, 16(10) :5232–5244, 1996.

[VBDBB+99] W. Vanden Berghe, K. De Bosscher, E. Boone, S. Plaisance, and G. Haegeman. The nuclear factor-kappab engages cbp/p300 and histone acetyltransferase activity for transcriptional activation of the interleukin-6 gene promoter. *J Biol Chem*, 274(45) :32091–8, 1999.

[WCBL02] W Wang, J. Cherry, D. Botstein, and H. Li. A systematic approach to reconstructing transcription networks in saccharomyces cerevisiae. In *Proceedings of the National Academy of Sciences of the United States of America*, volume 99, pages 16893–16898, 2002.

[Wei85] H. Weintraub. Assembly and propagation of repressed and derepressed chromosomal states. *Cell*, 42 :705–711, 1985.

[WHD+99] W. Wilbur, G. Hazard, G. Divita, J. Mork, A. Aronson, and A. Browne. Analysis of biomedical text for chemical names : a comparison of three methods. In *AMIA Symp*, pages 176–180, 1999.

[Win97] E. Wingender. Classification of eukaryotic transcription factors. *Mol Biol (Mosk)*, 31(4) :584–600, 1997.

[WLD+04] H. Wain, M. Lush, F. Ducluzeau, V. Khodiyar, and S. Povey. Genew : the human gene nomenclature database, 2004 updates. *Nucleic Acids Res.*, 32 :D255–7, 2004.

[Won01] L. Wong. Pies, a protein interaction extraction system. In *Pacific Symposium on Biocomputing*, volume 6, pages 520–531, 2001.

[WRR+88] Herr W., Sturm R., Clerc R., Corcoran L., Baltimore D., Sharp P., Ingraham H., Rosenfeld M., Finney M., and Ruvkun G. The pou domain : a large conserved region in the mammalian pit-1, oct-1, oct-2, and caenorhabditis elegans unc-86 gene products. *Genes Dev.*, 2(12A) :1513–1516, Dec 1988.

[WSC04] T. Wattarujeekrit, P. Shah, and N. Collier. Pasbio : predicate-argument structures for event extraction in molecular biology. *BMC Bioinformatics*, 5 :155+, 2004.

[XRS+00] I. Xenarios, D. Rice, L. Salwinski, M. Baron, E. Marcotte, and D. Eisenberg. Dip : the database of interacting proteins. *Nucleic Acids Res*, 28(1) :289–291, January 2000.

[Yak06] A. Yakushiji. *Relation Information Extraction Using Deep Syntactic Analysis*. PhD thesis, University of Tokyo, 2006.

[YDHS04] F. Yamout, R. Demachkieh, G. Hamdan, and R. Sabra. Further enhancement to the the porter's stemming algorithm. In *Workshop on Text-based Information Retrieval (TIR-04)*, 2004.

[YHRW02] H. Yu, V. Hatzivassiloglou, A. Rzhetsky, and W. Wilbur. Automatically identifying gene/protein terms in medline abstracts. *J. of Biomedical Informatics*, 35(5/6) :322–330, 2002.

[YM02] M. Yandell and W. Majoros. Genomics and natural language processing. *Nat Rev Genet*, 3(8) :601–610, 2002.

[YMTT05] A. Yakushiji, Y. Miyao, Y. Tateisi, and J. Tsujii. Biomedical information extraction with predicate-argument structure patterns. In *First International Symposium on Semantic Mining in Biomedicine*, pages 60–69, 2005.

[YPTM06] H. Yu, A. Paccanaro, V. Trifonov, and G. Mark. Predicting interactions in protein networks by completing defective cliques. *Bioinformatics*, 22(7) :823–829, April 2006.

[YTMT01] A. Yakushiji, Y. Tateisi, Y. Miyao, and J. Tsujii. Event extraction from biomedical papers using a full parser. *Pac Symp Biocomput*, pages 408–419, 2001.

[ZS04] G. Zhou and J. Su. Exploring deep knowledge resources in biomedical name recognition. In *Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, 2004.