UNIVERSITÉ D'AIX-MARSEILLE III

EXPÉRIENCE DE COUPLAGE ENTRE BASES DE DONNÉES FACTUELLES ET BASES DE DONNÉES BIBLIOGRAPHIQUES : IDENTIFICATION DANS MEDLINE DES GÈNES DÉCRITS DANS FLYBASE ET APPLICATION À L'EXTRACTION D'INFORMATIONS SUR LES INTERACTIONS GÉNÉTIQUES OU MOLÉCULAIRES À PARTIR DE PUBLICATIONS

THÈSE

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE DROIT, D'ÉCONOMIE ET DES SCIENCES D'AIX-MARSEILLE

en

SCIENCE DE L'INFORMATION ET DE LA COMMUNICATION

présentée et soutenue publiquement

le 4 Janvier 2001

par

Monsieur Ambroise Ingold

Sous la direction de

Monsieur Bernard Jacq

Monsieur Luc Quoniam

JURY

M. Bernard Jacq, Chargé de recherche

M. Henri Dou, Professeur

M. Luc Quoniam, Professeur

M. Philippe Dessen, Directeur de recherche (rapporteur)

M. Pierre Zweigenbaum, Titulaire d'une habilitation à diriger des recherches (rapporteur)

M. Xavier Polanco, Docteur d'état (rapporteur)

S O M M A I R E

REMERCIEMENT

Je voudrais tout d'abord remercier Henri Dou et Luc Quoniam pour la bourse de thèse dont j'ai bénéficié grâce à eux.

Je remercie Luc Quoniam pour avoir lancé le projet avec l'énergie qui le caractérise et d'avoir coordonné le travail d'équipe avec Violaine Pillet. Il m'a mis le pied à l'étrier et m'a encouragé.

Je remercie Bernard Jacq pour la curiosité et l'ouverture d'esprit dont il a fait preuve. Je le remercie aussi pour le temps qu'il a consacré au suivi de mon travail et pour sa participation à l'annotation des textes.

Je suis reconnaissant à Henri Dou d'avoir apporté son soutien aux moments importants, me permettant de mener ma thèse à bonne fin.

Je remercie mon entourage qui m'a soutenu et conseillé pendant les périodes d'incertitude et en particulier :

- mon père pour ses corrections,
- Alice pour ses relectures averties,
- Jean-Baptiste pour avoir souvent témoigné de l'intérêt pour la progression de mon travail,
- Laurence pour la finesses de son jugement.

Je remercie Valérie Leveillé, Marie Thérèse Maunoury, François Radvanyi et Marianne Tuefferd pour avoir corrigé mon manuscrit.

Je remercie Christian Jacquemin, William Turner et François Rechenmann pour avoir lu mon manuscrit et m'avoir donné des conseils.

Je remercie enfin les membres du jury pour leurs participations et leurs conseils.

INTRODUCTION

L'achèvement du projet génome ouvre la voie vers de nouvelles perspectives. Il s'agit d'exploiter les données acquises pour parvenir à comprendre le fonctionnement du vivant. De grandes bases de données capitalisent et organisent le savoir accumulé. Chacune répond à une problématique différente et la synthèse appartient finalement au chercheur lui-même. Les bases de données bibliographiques permettent l'accès au texte, qu'il s'agisse d'un résumé de publication, ou qu'il s'agisse du texte intégral de la publication. C'est là que se trouve l'information la plus complète, la plus détaillée, et la plus à jour. Les encyclopédies électroniques offrent un point de vue synthétique sur l'état du savoir. Les bases de données de résultats d'expériences permettent de formuler des hypothèses fécondes. Le va-et-vient entre les bases de données bibliographiques et les bases de données factuelles est rendu possible par des liens croisés. Pour les bases de données factuelles, il s'agit de maintenir des pointeurs vers de la bibliographie. Pour les bases de données bibliographiques, il s'agit, d'une part, de proposer des liens vers les résultats d'expériences, et d'autre part, d'indexer les textes à l'aide des nomenclatures maintenues par les encyclopédies électroniques. La masse des données en jeu est énorme. La vitesse avec laquelle le savoir s'accumule et s'actualise est grande. L'établissement de liens entre bases de données bibliographiques et bases de données factuelles ne peut plus être effectué manuellement. Comment l'automatiser ? Nous prenons deux exemples d'applications complémentaires pour y répondre.

Dans les publications scientifiques, qu'est-ce que nous aimerions voir repéré et lié à des données factuelles ? D'une part les objets, et d'autre part, les relations entre ces objets. Nous prendrons un exemple d'application pour chaque cas. Pour les objets, nous prendrons comme exemple les gènes de la Drosophile. Pour les relations, nous prendrons comme exemple les interactions génétiques ou moléculaires chez ce même organisme. Il s'agit d'un type de relation pertinent en génétique. La question est de savoir quand deux gènes collaborent dans un processus dans lequel ils sont impliqués. Dans certains cas (interaction moléculaire), cette collaboration s'explique par un contact physique entre des molécules. Plus généralement (interaction génétique), les mécanisme moléculaire reste inconnus ou l'interaction résulte de plusieurs interactions moléculaires s'enchaînant en cascade.

Notre travail comprend donc deux volets. Le premier volet consiste à repérer dans la base de données bibliographique *Medline*[1] les gènes qui sont répertoriés dans l'encyclopédie électronique sur la Drosophile *Flybase*. Le deuxième volet consiste à construire une base de données sur les interactions à partir des résumés de publications contenus dans *Medline*. Enfin, nous présentons le travail d'annotation permettant d'évaluer les résultats.

1. Le repérage des gènes et des interactions géniques

À quelles disciplines scientifiques pouvons-nous faire appel ? Il s'agit tout d'abord de recherche d'informations. Quels sont les textes qui citent tel ou tel gène ? Quel sont les textes qui traitent d'interactions génétiques ou moléculaires ? L'intelligence artificielle est concernée, elle aussi. Il s'agit de compréhension du langage naturel. L'analyse grammaticale des phrases permet de repérer des syntagmes qui sont éventuellement des noms de gènes.

[1] Les termes de l'index sont en italique. A l'endroit du textes où ils sont définis, ils sont en gras.

Elle permet aussi de repérer des relations qui sont énoncées dans les textes. La *bibliométrie* et la statistique textuelle sont concernées. Il s'agit de valoriser des textes de publication scientifique, de trouver le vocabulaire spécifique des interactions et d'amener le lecteur à découvrir des relations entre les gènes par l'étude de la cooccurrence. Nous discuterons des apports possibles de ces disciplines à travers l'étude de travaux comparables. Nous emprunterons à chacune d'elles des idées, des techniques et des moyens d'évaluation.

Nous pensons que l'accès à l'information textuelle doit se faire par des mots clefs. Pour l'identification des gènes, il s'agit d'utiliser les divers noms du gène ou de ses produits. Pour la reconnaissance des interactions, il s'agit d'analyser le vocabulaire présent pour détecter le thème de l'interaction. C'est d'ailleurs de cette façon que les utilisateurs des bases de données recherchent des informations sur les objets qui les intéressent ou sur les relations qu'entretiennent ces objets entre eux. Ils utilisent des mots clefs qui représentent, soit les objets, soit les relations.

Identifier les gènes cités dans un texte est une tâche difficile à automatiser car la nomenclature est complexe et l'usage ne suit pas toujours la norme. Des abréviations sont utilisées, mais il peut en exister plusieurs. La terminologie évolue avec la progression de la connaissance. Les anciennes dénominations vont former des alias. En outre, un gène peut être désigné par ses produits, en particulier les protéines dont il commande la synthèse. Des variations orthographiques existent, notamment avec la coupure des mots, l'usage optionnel des tirets ou des majuscules. Des contradictions finissent par apparaître : deux termes peuvent désigner le même gène. Le contexte peut primer quand il existe plusieurs indices concordants de la présence d'un même gène. Certaines interprétations devront être privilégiées à contexte équivalent. Il existe aussi des termes vagues, qui ne désignent pas un gène précis mais peuvent renvoyer à toute une collection de gènes.

Flybase rend compte de la variété de tel ou tel nom de gène mais pas du choix des dénominations. S'y retrouve mêlées des informations relatives à la norme, à l'historique, aux mécanismes biologiques (avec les noms de protéines), aux variations orthographiques, aux écarts acceptables par rapport à la norme, à l'usage ponctuel dans une publication, à l'usage fautif, au manque de précisions, etc. *Flybase* présente un inventaire. Elle n'explique pas comment identifier les gènes dans un texte.

La polysémie est présente dans le domaine des gènes de la drosophile. D'une part, les termes utilisés peuvent désigner tout autre chose que des gènes. Par exemple : *labial*, *blood* et *arrest* sont des noms de gènes. De même *N*, *h* et *if* sont des noms abrégés des gènes *Notch, hairy* et *inflated*. D'autre part, les noms de gènes peuvent rentrer dans la composition de terme désignant d'autres gènes ou tout autre chose. Par exemple, les noms de gènes *lethal of scute* et *Suppressor of Hairless* sont formés sur les noms de gènes *scute* et *Hairless*. Autre exemple, le nom de gène *scute* entre dans la composition de *Achaete-scute Complex* qui est un complexe de gènes ; *Polycomb* entre dans la composition de *Polycomb group* qui est un groupe de gènes. Ce phénomène d'appariement de plusieurs termes pour former une entité de sens est appelé *collocation*.

Le contexte permet souvent de lever les ambiguïtés. Il permet soit d'identifier une collocation, soit de donner plusieurs indices concordants de la présence d'un même gène.

Pour parvenir à automatiser l'identification des gènes, il s'agit de structurer correctement les connaissances nécessaires à l'interprétation et de trouver un algorithme efficace. L'organisation des données a été conduite grâce à une base de données relationnelle.

L'algorithme permettant l'interprétation du texte a été mis en œuvre grâce à des automatismes se succédant dans un ordre déterminé.

Plus qu'un algorithme d'identification des gènes, nous proposons une méthode pour corriger, structurer et enrichir des données déjà existantes sur la terminologie, de façon à permettre leur utilisation lors d'un processus automatique d'identification des gènes. Cette méthode passe par une confrontation des données présentes dans la nomenclature et des textes à annoter. Cette confrontation permet à la fois de mesurer l'efficacité du processus, de comprendre les problèmes, de corriger, annoter ou enrichir la nomenclature. La question est aussi d'obtenir des informations sur l'usage réel qui est fait de la nomenclature. Par exemple quelle est la fréquence d'utilisation des abréviations, des alias, etc. ?

La nomenclature maintenue par *Flybase* n'est pas tout à fait complète. Des variations orthographiques ont été omises. Nous proposons un système pour anticiper certaines de ces variations orthographiques. Il s'agit d'enrichir la nomenclature par analyse automatique des textes, de façon à valider les variations orthographiques anticipées.

2. *Extraction d'information sur les interactions génétiques ou moléculaires*

Le deuxième volet de notre travail consiste à construire une base de données sur les interactions génétiques ou moléculaires à partir de résumés de publications. Il s'agit d'une tâche d'extraction de connaissances à partir de texte. Nous utilisons principalement la cooccurrence pour y parvenir. Il s'agit de repérer les couples de gènes cités dans au moins une phrase. Nous utilisons en complément un indicateur statistique évaluant la probabilité qu'une phrase décrive une interaction. Cet indicateur est basé sur l'existence d'un vocabulaire spécifique aux interactions génétiques ou moléculaires. Il s'agit de résultats obtenus au CRRM par Violaine PILLET et que nous avons intégrés à notre système (2000).

Les travaux sur l'extraction d'information sur les interactions génétiques ou moléculaires sont de deux types. Dans un premier type d'étude, des matrices de cooccurrence sont calculées pour savoir si l'apparition d'un gène est corrélée avec l'apparition d'un autre gène. Ces études permettent de découvrir des relations fonctionnelles évidentes ou cachées et apportent donc une aide à la découverte. Les résultats sont évalués sur la base de leur utilité pour le chercheur. Il est difficile de savoir dans quelle mesure une information initialement présente dans les textes est ou n'est pas retenue.

Dans un autre type d'étude, des modèles d'énoncés d'interactions sont inventoriés et les motifs textuels correspondants sont recherchés dans les textes. Cependant les modèles utilisés sont simples et les motifs sont donc très spécifiques. Par exemple, il s'agira de rechercher des verbes d'action comme *bind* et de les associer aux syntagmes nominaux voisins, qui sont en principe des noms de gènes ou de protéines. La spécificité des motifs recherchés va assurer une grande qualité des réponses fournies. La précision sera donc bonne. Inversement la quantité d'information extraite sera relativement faible car beaucoup de descriptions d'interactions ne correspondront pas au modèle prédéfini. Le rappel sera donc mauvais.

3. *Les bénéfices de l'annotation*

Dans la plupart des études, la perte d'information, due à la trop grande spécificité des motifs textuels recherchés ou plus généralement à la sélection qui est opérée sur les textes, n'est pas évaluée. En effet, dans ces études, l'évaluation n'est menée que sur les textes qui présentent déjà des caractéristiques bien précises. Nous pensons que l'annotation des textes

doit se faire avant toute sélection. C'est la seule façon de connaître la quantité d'informations qui est perdue après la sélection.

En outre, cette annotation est très instructive. Elle permet de se rendre compte du fait que les interactions sont souvent décrites dans des énoncés très complexes. Ces énoncés se prêtent difficilement à des recherches de motifs textuels précis. Nous proposons donc un système basé sur la recherche de termes simples. Ces termes sont choisis pour les renseignements qu'ils apportent à eux seuls – c'est à dire en dehors de toute combinaison – sur la présence d'une interaction. Ceci est apprécié par une corrélation statistique entre leur utilisation dans une phrase et la présence d'une interaction dans cette même phrase. À chacun de ces termes est associé un coefficient qui a été calculé sur un échantillon d'apprentissage. La moyenne des nombres ainsi trouvée dans une même phrase nous renseigne sur la probabilité d'avoir une ou plusieurs interactions dans la phrase. Les phrases dont le vocabulaire est considéré comme favorable seront annotées. Cette annotation consiste à repérer tous les couples de gènes en présence.

Le document est organisé en deux parties. La première partie donne les éléments nécessaires à la compréhension en ce qui concerne la veille technologique, la bibliométrie, la biologie et les techniques de recherche et d'extraction d'informations. L'analyse critique des travaux menés dans ce domaine est fournie dans cette partie. Nous décrivons ensuite le programme de recherche auquel nous prenons part, ainsi que les principaux résultats sur lesquels nous nous appuyons. La deuxième partie présente les réalisations effectuées, elle fournit les résultats obtenus et donne des méthodes d'évaluation de ces résultats.

Partie 1

État de l'Art

Chapitre 1 Contexte scientifique de l'étude

I. VEILLE TECHNOLOGIQUE, INTELLIGENCE ÉCONOMIQUE ET ANALYSE DE L'INFORMATION TEXTUELLE

Je travaille au ***CRRM***[2] qui est un laboratoire dont l'objet de recherche est la ***veille technologique***. La veille technologique est une discipline orientée vers l'entreprise. Elle permet aux décideurs de mieux comprendre l'environnement scientifique et technique de façon à orienter leurs choix stratégiques (DOU, 1995).

La veille technologique fait partie d'une démarche plus globale d'observation de la concurrence, des marchées, de la législation, de la réglementation, des normes, des évolutions sociales, etc. On parle alors d'***intelligence économique*** (MARTINET, 1995).

Les informations sont collectées, analysées et synthétisées par des professionnels de la veille avant d'être transmises à la direction de l'entreprise (JAKOBIAK, 1998). Le veilleur s'intéresse aux informations formelles, comme aux informations informelles. Les premières regroupent les publications scientifiques, articles de presse, rapports, études, notices bibliographiques, rapports de dépôts de brevets, bases de données, etc. Les secondes regroupent les comptes rendus de visites dans les salons professionnels, les rumeurs, les informations obtenues auprès des fournisseurs ou clients, etc. Les informations collectées sont le plus souvent non confidentielles. Dans tous les cas, elles ont été obtenues légalement.

Pour l'information scientifique et technique, les sources d'informations utilisées sont pour l'essentiel présentes dans des bases de données. Les données en jeu sont très nombreuses, elles se prêtent donc particulièrement bien aux études globales. La veille technologique permet de déterminer quels sont les thèmes de recherche les plus en vogue, de connaître les domaines techniques dans lesquels les concurrents déposent leurs brevets, de réaliser des réseaux de co-auteurs, voire d'anticiper des tendances futures à partir de signaux faibles.

Les techniques utilisées sont issues de la ***bibliométrie***. Dans cette discipline, il s'agit d'effectuer des mesures (dénombrements ou études statistiques) sur la science et les techniques à partir de publications. Le facteur d'impact (*impact factor*) que calcule l'*ISI*[3] pour évaluer l'importance d'une revue dans sa discipline est un très bon exemple d'étude bibliométrique (MAGRI, 1997). Quand les études servent au pilotage de la politique de recherche on parle de ***scientométrie*** (BARRÉ, 1995). Quand les études servent à optimiser le fonctionnement des bibliothèques, par exemple la souscription d'abonnement à des revues, on parle de ***bibliothéconomie***.

La bibliométrie permet d'étudier les producteurs (chercheur, équipe, laboratoire, entreprise, pays, …) ou les diffuseurs (éditeur, périodique, colloque,…) d'un point de vue quantitatif ou qualitatif. Elle permet aussi de cartographier un domaine de recherche. Des graphes de co-auteurs peuvent être réalisés, des mots clefs peuvent être identifiés.

---

[2] Centre de recherche rétrospective de Marseille. http://crrm.u-3mrs.fr

[3] Institute for Scientific Information. http://www.isinet.com/isi/

L'étude que nous proposons sur les interactions entre les gènes fait bien partie du champ de la bibliométrie. Il s'agit d'exploiter des ressources bibliographiques existantes pour obtenir une vision synthétique d'un domaine de recherche donné. Il y a d'ailleurs une forte analogie entre les réseaux de co-auteurs et les réseaux de gènes en interactions. Dans les deux cas, il s'agit d'offrir une vision synthétique des collaborations qui interviennent entre différents acteurs dans la réalisation d'une certaine tâche.

Dans la plupart des cas, les études bibliométriques ne s'intéressent pas au champ résumé. Les champs utilisés sont les champs auteur, affiliation, date, source (nom du journal par exemple), mot clef, code de classement, etc. Ces champs sont appelés ***champs contrôlés***. Ces champs contiennent une information de nature très différente de celle qui est contenue dans le champ résumé. Les modalités possibles pour les champs contrôlés sont assez limitées. Il s'agit d'un mot, d'un code, d'une date. En revanche le champ résumé contient du texte rédigé. Dans le domaine du traitement d'enquête, on appelle cela du ***texte libre*** : la personne interrogée rédige librement sa réponse. Cela correspond à une réponse à une question ouverte : que faites-vous ? Tandis que les champs de description du document correspondent chacun à une réponse à une question fermée : qui a participé à l'écriture du document ? Où a-t-il été publié ? Etc. Dans le domaine de l'intelligence artificielle on appelle le texte rédigé du ***langage naturel***, par opposition aux langages informatiques qui manipulent des symboles, des nombres, des équations et des instructions.

La nature des champs est très différente ; les moyens d'analyse seront donc différents. Alors que pour le champ auteur, il suffit d'extraire les associations de noms présents pour obtenir un graphe de co-auteurs, il n'en va pas de même pour réaliser un graphe de co-citation des gènes. Cela demande des traitements préliminaires : reconnaître les noms de gènes en présence dans le texte et les associer au gène qu'il désigne (à travers un numéro unique propre à chaque gène). Ce problème, l'identification des gènes dans les textes, est assez complexe comme nous le verrons dans la section Partie 2Chapitre 1I[4]. Le traitement a nécessité l'utilisation d'un dictionnaire des gènes. Ce dictionnaire décrit la terminologie employée pour les gènes de la drosophile. Le traitement a aussi nécessité l'emploi d'une base de données relationnelle. Cette base était en effet indispensable pour organiser correctement le dictionnaire. Elle a aussi servi à mémoriser les annotations faites par l'expert et par le programme. Elle permet de faire toutes sortes de comparaisons entre les données, de façon à affiner la terminologie des gènes et à résoudre leurs identifications dans les textes.

II. DE LA GÉNÉTIQUE À LA BIOINFORMATIQUE

A. LA GÉNÉTIQUE

1. Notions de base

Le patrimoine héréditaire d'un être vivant se transmet de génération en génération. L'ensemble de l'information génétique commune aux individus d'une même espèce constitue le ***génome***. Cette information est représentée dans plusieurs macros molécules d'***ADN*** (***Acides DésoxyriboNucléiques***). L'*ADN* est constitué d'une succession de molécules appelées ***nucléotides***. C'est la ***séquence***, c'est à dire l'ordre dans lequel ces éléments de bases sont assemblés dans la macro molécule qui constitue l'information.

[4] Pour consulter les renvois, on pourra se référer au plan détaillé.

La molécule d'*ADN* interprétée pour donner naissance à des protéines. On peut dire que la molécule d'*ADN* contient le plan de fabrication des protéines.

Une ***protéine*** est, comme la molécule d'*ADN*, constituée d'une succession d'éléments, les ***acides aminés***. L'ordre dans lequel les nucléotides sont disposés dans l'*ADN* va déterminer l'ordre dans lequel les acides aminés seront disposés dans la protéine synthétisée. La règle qui permet de passer de l'un à l'autre est identique (ou presque) pour tout le vivant, c'est le ***code génétique***.

Les protéines sont très importantes dans le fonctionnement du vivant. Les enzymes, qui sont des catalyseurs naturel pour les réaction chimiques qui ont lieux au seins du vivant, sont le plus souvent des protéines. Les protéines interviennent dans les mécanismes de régulation et de transport de signal au seins de la cellule ou entre les cellules. On parle à ce propos de ***voie de régulation*** ou de ***voie de signalisation***. La capacité d'une protéine à réaliser une fonction est fortement liée à sa structure spatiale. La conformation dans l'espace d'une protéine est décrite en terme de ***structure*** dans laquelle on peut reconnaître des motifs (HUNTER *et alii*, 1993).

Le patrimoine génétique d'un individu est décomposé en unités d'informations appelées ***gènes***. Il y a environ 20000 gènes pour la drosophile et 30000 pour l'homme. Dans les cas simples, un gène code pour une protéine. Il y a débat pour définir exactement ce qu'est un gène (WAIN *et alii*, 2000). Nous adopterons implicitement la définition que se donne *Flybase* en utilisant les informations issues de cette base de données.

La synthèse des protéines se fait en deux étapes : la ***transcription*** et la ***traduction***.

La portion de l'*ADN* correspondant à un gène est tout d'abord recopiée presque à l'identique, on dit ***transcrite***, dans une molécule messagère appelée ***Acide Ribo Nucléique messager*** ou ***ARN***$_m$. Dans certain cas, la molécule transcrite subit des transformations avant d'être traduite ; c'est la maturation. Cette opération se fait souvent par coupure puis recollement de certains segments ; c'est l'***épissage***. La molécule issue de la transcription mais non encore parvenue à maturité est appelée ***ARN précurseur***. L'ARN_m est ensuite *traduit* en protéine selon les spécifications du code génétique. Cette opération est appelée ***traduction***.

Tous les gènes ne sont pas actifs en même temps. Leur expression est contrôlée à plusieurs niveaux.

Un premier contrôle de l'expression est effectué au niveau de la *transcription*. Certaines protéines vont pouvoir se fixer sur *l'ADN* pour empêcher ou au contraire favoriser l'expression d'un gène situé à proximité.

Un deuxième contrôle de l'expression est effectué au moment de la maturation. Ce contrôle de l'expression génétique est dit ***post-transcriptionnel***.

Après la traduction, la protéine produite va elle-même pouvoir subir des transformations qui vont par exemple activer sa fonction. C'est le contrôle ***post-traductionnel***.

Ces actions au niveau moléculaire ont souvent des conséquences observables facilement sur l'individu. Typiquement, une déficience génétique va se traduire par une modification voire une malformation de l'individu. On parle alors d'individu ***mutant***. L'individu originel est qualifié de ***sauvage***. Des défauts génétiques distincts peuvent avoir des conséquences

similaires. On fait donc la distinction entre ***phénotype*** et ***génotype***. Le phénotype correspond à l'apparence de l'individu tandis que le génotype correspond à sa constitution génétique.

Les deux composantes d'un même gène sont désignées par le terme d'***allèles***. Des allèles distincts peuvent conduire à des individus tous sains mais ayant des caractéristiques différentes telles que la couleur des yeux. Dans ce cas, la distinction sauvage-mutant n'a plus de sens et n'est pas utilisée.

L'existence simultanée dans une population de plusieurs allèles d'un même gène est appelée le ***polymorphisme***. Son étude est intéressante car elle donne accès aux mécanismes dans lesquels le gène est impliqué, autrement dit à sa fonction.

2. Définition des interactions

Une ***interaction*** peut s'exercer de multiples façons. Nous prendrons une définition très large de la notion d'interaction de façon à récolter un maximum d'informations à extraire dans les textes que nous avons annotés manuellement pour les besoins de l'expérience. Un développement possible de notre travail serait de faire des distinctions entre chaque type d'interaction.

Il existe deux grandes catégories d'interactions : interactions moléculaires et interactions génétiques.

Les ***interactions moléculaires*** correspondent à un contact entres deux molécules. Les partenaires de ces interactions sont des protéines ou des acides nucléiques (*ADN* ou *ARN*), mais un partenaire au moins est une protéine. Il existe donc trois cas de figures :

- Interaction protéine – *ADN*,
- Interaction protéine – *ARN*,
- Interaction protéine – protéine.

Cela correspond par exemple au cas d'une protéine qui va se fixer à une séquence d'*ADN* spécifique et activer la transcription d'un gène. Autres exemples : une enzyme va couper une molécule d'*ARN* lors de la maturation de celle-ci, ou encore, deux protéines vont s'assembler pour former un complexe.

Dans tous les cas, les phénomènes moléculaires vont avoir des conséquences au niveau génétique. Autrement dit, des expériences de génétiques (par exemple l'observation du phénotype des individus obtenu par croisement) vont trahir les phénomènes moléculaires sous-jacent. Par exemple, il se peut qu'un gène soit inhibé par un autre gène. Ceci nous amène à définir la notion d'interaction génétique.

Les ***interactions génétiques*** correspondent à des modifications dans l'action d'un gène induites par l'expression d'un autre gène. Typiquement, les interactions génétiques sont mises en évidence par une observation des *phénotypes*. Par exemple, si le phénotype d'un mutant sur un premier gène est aggravé ou au contraire sauvé par une mutation sur un deuxième gène, alors il y a interaction entre les deux gènes.

Une interaction génétique peut être la conséquence directe d'une interaction moléculaire mais il se peut aussi qu'elle soit la conséquence d'une cascade d'interactions moléculaires. Les interactions génétiques sont donc des interactions dont on ne connaît pas le mécanisme moléculaire ou qui sont la conséquence de plusieurs interactions moléculaires.

En définitive, la notion d'interaction que nous prenons en compte recouvre des réalités biologiques variées :

- Possibilité de fixation d'une protéine sur *l'ADN*
- Régulation post-transcriptionnelle
- Modification post-traductionnelle
- Formation de complexe protéique
- Activation ou inhibition d'un gène par un autre
- Participation à des voies de signalisations
- etc.

B. LE PROJET GÉNOME

Depuis la découverte de l'importance des molécules d'*ADN* dans le stockage de l'information génétique, les biologistes ont formé le projet d'en connaître la séquence complète. Des actions de séquençages du génome complet de plusieurs organismes ont vu le jours. Le plus grand chantier, en terme de moyen mis en œuvre, est celui qui concerne l'homme ; c'est ce que l'on a appelé ***le projet génome***.

Actuellement, plus de trente-cinq génomes bactériens sont séquencés. On compte aussi cinq génomes d'organismes plus évolués (des eucaryotes), à savoir, la levure *Sacharomyces cerevisæ*, le ver *Cænorhabditi elegans*, la mouche ***Drosophila melanogaster***, la plante *Arabidopsis thaliana* et enfin l'Homme. Ce dernier n'est encore qu'un « premier jet » mais renferme déjà une quantité considérable d'informations nouvelles (THE GENOME SEQUENCING CONSORTIUM, 2001).

Avec l'apparition de grandes quantités d'informations numérisées disponibles, est apparue une nouvelle discipline scientifique : la ***bioinformatique***. Cette discipline utilise des connaissances issues de la biologie, de l'informatique, des mathématiques et notamment des statistiques. ANDRADE et SANDER (1997) définissent la bioinformatique comme un nouveau domaine de recherche qui, à partir de données biologiques et par l'utilisation de méthodes informatiques, permet de créer des connaissances nouvelles dans le domaine de la biologie elle-même.

Quand il s'agit d'exploiter des informations sur le génome on parle alors de ***génomique***. Ce terme a été inventé par THOMAS H. RODERICK *et alii* en 1986 lors d'une discussion sur le nom d'un nouveau journal. Ce journal, *Genomics*, avait pour objet les données de séquences, la découverte de nouveau gènes, la cartographie génétique et plus généralement les nouvelles techniques en génétique. Ces études de génétique se consacrent au génome pris comme un tout, contrairement par exemple à l'analyse d'un ou de quelques gènes impliqués dans tel ou tel mécanisme biologique.

Au départ, la génomique s'est consacrée principalement à l'analyse des données de séquences. La détermination dans le génome des séquences codantes, à savoir celles qui correspondent à des gènes ou à des séquences régulatrices, est un problème classique de la génomique.

Elle s'oriente désormais vers l'étude de la fonction des gènes. Le terme de ***génomique fonctionnelle*** a été formé pour désigner cette nouvelle tendance (HIETER *et alii*, 1997).

La relation entre la séquence et la fonction est une chose très complexe. Il est difficile de prévoir la forme d'une protéine à partir de sa seule séquence en acides aminés. Or la

structure spatiale est essentielle dans la détermination de la fonction de la molécule. Ainsi, même quand on se place dans le schéma simplifié qui affirme que l'on peut lire la protéine dans la séquence génétique, on voit que l'on ne peut pas prévoir la fonction d'un gène à partir de l'étude de sa seule séquence. C'est une des gageures de la bioinformatique que de parvenir à comprendre le passage entre séquence génétique et structure protéique pour enfin accéder à la fonction (ATTWOOD *et alii*, 2000).

C. LA POST-GÉNOMIQUE

Le grand projet de séquençage des génomes n'apparaît que comme une étape dans la compréhension des phénomènes biologiques. Le terme de ***post-génomique*** a été formé pour désigner cette évolution. KANEHISA *et alii* (2000), proposent d'employer le terme de génomique (*genome informatics*) pour désigner l'utilisation de l'informatique pour gérer les grandes quantités d'informations issues de l'étude des génomes, alors que la post-génomique (*post genome informatics*) aura pour but d'arriver à comprendre les phénomènes biologiques sous-jacents à partir de l'analyse informatique des données issues de l'analyse des génomes.

Pour analyser la fonction des gènes à l'échelle du génome, il faut être capable de comprendre les relations complexes qu'entretiennent les mondes des *ADN*, des *ARN* et des protéines. Ne s'intéresser qu'au génome n'est pas suffisant. Les concepts de ***protéome*** et de ***transcriptome*** ont été forgés à cet effet. WILKINS a créé le terme de protéome en 1994 et l'a défini comme le complément protéique exprimé par le génome (WASINGER, 1995). De la même façon, le transcriptome se défini comme l'ensemble des transcrits à l'échelle du génome.

La ***protéomique*** est l'étude de l'ensemble des protéines exprimées dans une cellule à un instant donné dans le but d'obtenir une vision globale des processus cellulaires (ROCHA *et alii*, 2000). Il s'agit en particulier de savoir avec quel substrat (*ARN* , *ADN*, autre protéine, autre molécule) les protéines exprimées interagissent. La *protéomique* est un domaine de recherche très actif, mais le but poursuivi est ambitieux car les expériences sont difficiles à mettre en œuvre.

La ***protéomique structurale*** se définit comme l'étude de la forme des protéines. Cependant, le terme de ***génomique structurale*** est aussi employé. Comme nous l'avons déjà signalé, il est difficile de prévoir la structure d'une protéine à partir de la séquence des acides aminés qui la compose. L'étude expérimentale de la structure des protéines à l'échelle du génome est donc un travail nécessaire mais très ambitieux (WILKINS *et alii*, 1996).

Pour mieux comprendre les relations entre génome et transcriptome, des études expérimentales à grande échelle sont entreprises pour produire des ***données d'expressions***. On cherche par ces expériences à réunir de grandes quantités d'informations sur les ARN_m transcrits. Ces informations reflètent directement l'activité d'expression des gènes selon le tissu observé, le stade du développement, l'état normal ou pathologique des cellules, etc. La technologie des ***puces à ADN*** est utilisée (The chipping forecast, 1999). La 'puce' est constituée d'un support sur lequel sont greffées plusieurs milliers à plusieurs dizaines de milliers de sondes. Chaque sonde va être capable de reconnaître spécifiquement une substance telle qu'une séquence donnée d'ARN_m. Les *puces à ADN* permettent donc de réaliser simultanément plusieurs dizaines de milliers d'expériences simultanément sur une toute petite surface. La logique est donc encore celle d'une production en masse de données, comme pour le séquençage.

Pour l'étude des interactions entre protéines, il existe aussi des techniques de production de données en masse. La méthode dite du ***double hybride*** (FIELDS *et alii* ,1989), l'***électrophorèse bidimensionnelle sur gel*** et la ***spectrométrie de masse*** peuvent par exemple être utilisées. La méthode du double hybride est utilisée à l'échelle du protéome pour la levure.

La *post génomique* se donne pour objectif d'intégrer toutes ces données quantitatives d'expressions des gènes (CHEE *et alii*, 1996). Il s'agit par exemple d'intégrer des données d'expression avec des informations sur les voies métaboliques (NAKAO *et alii*, 1999), le but étant principalement de prévoir la fonctions des protéines (GERSTEIN *et alii*, 2000).

Les informations de séquence, même si elles ne suffisent pas, sont très utilisées pour déterminer la fonction des gènes comme nous allons le voir dans la section suivante.

D. UTILISATION DU PROJET GÉNOME POUR ACCÉDER À LA FONCTION DES GÈNES

En premier lieu, notons que la notion de fonction est difficile à définir. Dans les bases de données sont présentes toutes sortes d'informations qui touchent d'une façon ou d'une autre à la fonction des gènes. Il existe principalement quatre types de fonctions :

- Fonction moléculaire : dans quels processus biochimiques les produits du gène sont impliqués
- Fonction cellulaire : dans quel processus biologique le gène est impliqué
- Fonction dans le développement : à quel stade et dans la formation de quel organe le gène est impliqué
- Fonction d'adaptation : comment le gène participe à la compétitivité de l'organisme dans son milieu.

Ainsi, plusieurs niveaux de description coexistent et il est difficile de construire une notion de fonction qui va rassembler toutes les approches possibles (DAVIDSON et APWEILER, 1999).

Quoi qu'il en soit, les données de séquences sont très utiles pour accéder à la fonction des gènes.

La première façon d'utiliser les données de séquences est de faire des comparaisons avec des séquences dont la fonction est déjà connue. En effet, une ***similarité*** dans la séquence trahit souvent l'existence d'un gène ancestral commun. Dans ce cas, les gènes sont qualifiés d'***homologues***. Cette notion a été définit par FITCH en 1970 (FITCH *et alii*, 1970). La conséquence est que bien souvent on a une analogie de fonction. La recherche de similarité fait partie des méthodes de base de la bioinformatique (BORK *et alii*, 1998). Une séquence étant connue, il s'agit d'identifier les éventuelles séquences homologues déjà présentes dans les banques de données de séquences. Ces recherches de similarités se font le plus souvent par des méthodes tel que ***BLAST***[5] qui permettent de cribler les banques de séquence à la recherche de séquence homologue (ALTSCHUL *et alii*, 1990). La recherche d'homologie vaut aussi bien pour des gènes situés dans le même génome que pour des gènes appartenant à des génomes différents. C'est d'ailleurs ce qui justifie l'étude d'organismes aussi éloignés de l'homme que la drosophile. On parle à ce sujet d'***organismes modèles***.

[5] Accessible sur http://www.ncbi.nlm.nih.gov/BLAST/

La comparaison des séquences est une méthode si puissante qu'elle permet d'obtenir des informations sur les interactions génétiques sans avoir aucune connaissance préalable sur la fonction des gènes. En effet, les événements de fusions de gènes au cours de l'évolution trahissent souvent des interactions génétiques entre les gènes qui ont fusionné. Ainsi il est possible grâce une comparaison purement informatique du génome de différents organismes de détecter des interactions génétiques (ENRIGHT *et alii*, 1999).

Pour avoir une vision globale de l'information, le biologiste a besoin de consulter plusieurs bases de données, d'où l'intérêt des travaux sur la création de liens entre bases de données.

III. LA RECHERCHE ET L'EXTRACTION D'INFORMATIONS TEXTUELLES

Pour communiquer le résultat de leurs recherches, les scientifiques écrivent des articles qui sont publiés sur divers supports. Ces informations sont disponibles par l'accès au résumé que permettent les bases de données bibliographiques telles que *Medline*. Les bases de données factuelles fournissent, elles aussi, des informations sous forme de textes écrits directement par des annotateurs. Dans le domaine des interactions génétiques et moléculaires, on peut citer ***SWISS-PROT***[6] (BAIROCH *et alii*, 2000) qui est une base de données sur les protéines particulièrement riche en annotations. Dans cette base de données, de nombreuses informations sont données sur la structure et la fonction des protéines, ainsi que sur les modifications post-traductionnelles subies par les protéines.

Dans les deux cas, ces informations, de nature textuelle, sont inaccessibles à la compréhension directe de l'ordinateur. On dit qu'ils sont écrits en langage naturel, par opposition aux langages formels utilisés en informatique. Ainsi, par exemple, l'essentiel de l'information sur les interactions génétiques et moléculaires n'est accessible que par le texte.

Pour accéder à l'information contenue dans les textes, deux domaines de recherche peuvent être mis à contribution : la recherche d'information textuelle et l'extraction d'information textuelle. Nous allons dans cette partie définir ces domaines et donner des exemples de réalisation.

A. LA RECHERCHE D'INFORMATIONS TEXTUELLES

Les techniques de ***recherche d'informations textuelles*** (***RI***) sont directement issues de la recherche documentaire qui est une discipline ancienne, antérieure à l'apparition des ordinateurs. Le but poursuivi par ces techniques est de permettre un accès au document plus rapide que la consultation intégrale de la collection des documents. L'élaboration d'index est la technique la plus simple. Plus généralement, un système de représentation des documents est utilisé. La requête de l'utilisateur est représentée dans un autre système de représentation. Requête et document sont comparés par l'appariement de leurs représentations. L'ensemble des documents appariés est présenté à l'utilisateur avec éventuellement un indice de pertinence. Il existe deux modèles principaux de recherche documentaire : le modèle booléen et le modèle vectoriel.

Dans le premier cas, la requête s'exprime à travers une expression booléenne, par exemple : auteur=Salton ET (année=1980 OU année=1981). L'appariement ne se fait que s'il y a correspondance exacte, c'est à dire si les caractéristiques du document correspondent

[6] Accessible sur http://www.expasy.ch/sprot/

exactement à la requête. Ce système est très largement utilisé, aussi bien pour les bases de données bibliographiques que pour les moteurs de recherche sur internet.

Dans le cas du modèle vectoriel, on recherche une similitude entre document et requête plutôt qu'une correspondance exacte. Cette similitude est une quantité qui prend des valeurs entre zéro et un. Elle est d'autant plus grande que document et requête ont des mots en communs. Dans ce modèle, les documents et les requêtes sont représentés par des vecteurs dans un espace vectoriel. La similitude entre document et requête est calculée par le cosinus de l'angle que font les deux vecteurs.

Les méthodes de recherche documentaires doivent être évaluées sur la quantité et la qualité des réponses qu'elles fournissent. Deux indicateurs sont utilisés pour cela. Il s'agit du ***taux de rappel*** et du ***taux de précision*** (SALTON *et alii*, 1983). Le premier correspond à la proportion des documents trouvés (parmi les documents cherchés). Le second correspond à la proportion de documents pertinents (parmi les documents ramenés).

Pour améliorer la performance du système de recherche, des traitements sont nécessaires. Il s'agit de simplifier la représentation des documents afin d'éviter que des documents similaires soient considérés comme trop différents (FALOUTSOS *et alii*, 1995). Une des premières étapes consiste à éliminer les mots qui à eux seuls n'apportent pas d'information sur le document (VAN-RIJSBERGEN *et alii*, 1979). Ces mots sont appelés *mots vides*[7]. Ce pré-traitement a été utilisé lors de la détermination du vocabulaire spécifique de l'interaction génétique ou moléculaire. Une seconde étape dans le traitement classique des documents, consiste à lemmatiser les textes (SALTON, 1989). Il s'agit de faire disparaître les différences morphologiques, par exemple les marques de pluriel, de féminin ou de conjugaison. Toutes les formes fléchies sont ramenées à une forme unique qui est appelée le ***lemme***. Cette technique a été utilisée une première fois pour déterminer le vocabulaire spécifique et une seconde fois pour détecter les phrases qui par leur vocabulaire semblent décrire des interactions génétiques ou moléculaires.

D'autres directions de recherche existent en recherche d'information. Il y a notamment les techniques visant à organiser automatiquement les documents. Ces techniques permettent de faciliter la consultation et donc d'accéder plus facilement aux documents et à l'information. Cette organisation peut consister en des opérations de sélection de documents selon un thème, de tris des documents selon leur pertinence par rapport à une problématique, de classement en différentes rubriques, etc. A titre d'exemple, USUSAKA *et alii* proposent une méthode basée sur l'apprentissage de cas pour sélectionner des résumés traitant d'un thème particulier (1998).

Dans le domaine de la veille technologique et de l'intelligence économique, GOUJON propose un système d'analyse de texte qui met en évidence des segments de textes ayant des traits particuliers (2000). Cette technique permet, d'après son auteur, d'analyser le contenu d'un ensemble de documents tels que des brevets sans avoir à les lire intégralement.

Voyons maintenant dans quelle mesure notre travail s'inscrit dans la recherche d'informations et comment il s'en distingue.

[7] Les termes de l'index sont en italique. Là où ils sont définis, ils sont aussi en gras. La consultation de l'index permet de trouver la page où ils sont définis (numéro de page en gras) et les pages où ils sont utilisés.

Le premier volet de notre travail, à savoir, l'identification de gènes dans les textes, s'inscrit en partie dans la recherche d'informations. Il s'agit de repérer des objets pertinents dans des textes.

Cependant, nous faisons une distinction entre la détection d'une ***occurrence*** d'un gène et l'identification d'un gène. Dans le premier cas, il s'agit de détecter la présence d'une référence à un gène, autrement dit de repérer qu'un segment de texte est un nom de gène. Dans le second cas, il s'agit en plus d'associer le segment de texte à un gène bien précis d'une liste établie préalablement.

Quand nous voulons parler indifféremment de l'une ou de l'autre des tâches nous emploierons l'expression ***reconnaissance de gènes***. Dans cette expression, il n'est pas précisé s'il s'agit simplement de repérer la présence d'un gène ou s'il s'agit de déterminer de quel gène exactement on parle.

La tâche qui est la nôtre est bien celle de l'identification des gènes dans les textes. Elle se rapproche de la recherche d'informations. Cependant, dans la recherche classique d'informations, un seul objet est recherché, alors que nous allons rechercher tous les gènes de la drosophile. Nous classerons donc notre travail dans la création de liens entre bases de données textuelles et factuelles comme nous le verrons section IV.

B. L'EXTRACTION D'INFORMATIONS TEXTUELLES

La recherche d'informations, dans son expression la plus simple, consiste à extraire un document ou un segment de document, c'est à dire une portion de texte. Le résultat est destiné à la lecture humaine et non à une exploitation informatique. Si l'on cherche à retourner une information codée dans un langage accessible à l'ordinateur, on quitte le champ de la recherche d'informations pour entrer dans celui de l'***extraction d'informations***.

L'extraction d'informations peut être considérée comme une branche de l'informatique. Il s'agit de répondre à une question bien précise. La réponse devra être codée dans un format défini par avance (JACQUEMIN *et alii*, 2000). Des exemples typiques de tâches d'extraction d'informations sont donnés dans la série de conférences ***MUC*** (MUC-6, 1996). Les ***Message understanding conferences*** sont des compétitions organisées dans le domaine de l'extraction d'informations. Dans les épreuves, il s'agit par exemple de savoir quelles sont les entreprises qui fusionnent, se créent, passent des accords, etc. Pour réaliser une telle tâche, il faut être capable de réaliser des sous-tâches. La compétition est organisée en épreuve correspondant chacune à une sous-tâche. Je décris dans les sections suivantes chacune des sous-tâches définies par les conférences *MUC*.

- La ***reconnaissance d'entités nommées (REN)***

Il s'agit de reconnaître les entités tels que des noms d'entreprise, des noms de personnes, des noms de lieux, des dates, etc. Les méthodes utilisées peuvent être basées sur un apprentissage statistique d'exemples (BIKEL *et alii*, 1997), sur le repérage d'indices comme les titres honorifiques (Monsieur, Docteur, etc.), ou sur la recherche de patron syntaxique. Pour chaque entité rencontrée, une marque SGML (i.e. une chaîne de caractères) est posée de façon à délimiter la portion de texte repérée.

De tels travaux existent pour la reconnaissance des labels et nous y reviendrons dans la partie consacrée aux travaux concernant la reconnaissance de gènes sans lexiques (section Chapitre 2I.B).

- La ***résolution d'anaphore*** (*co-reference resolution*)

Il s'agit de repérer quand, dans un texte, il est fait référence plusieurs fois à une même entité, même si cette entité est nommée de façons différentes ou si un pronom personnel est utilisé. Cela recouvre en particulier la détection de lien d'abréviation, avec comme exemple d'utilisation la construction de dictionnaire d'acronymes à partir de corpus. Ce type de tâche est utile pour la reconnaissance des interactions car il est fréquent qu'un gène soit nommé de plusieurs façon dans le même résumé, précisément dans le cas d'utilisation d'abréviation. En revanche, le cas de l'utilisation d'un pronom pour faire référence à un gène semble être très peu fréquent dans notre corpus.

Il existe des travaux sur la construction, à partir de corpus, de dictionnaire d'acronymes pour les noms de gènes. Nous y reviendrons à la section Chapitre 2I.C.

- Le remplissage d'un formulaire simple (*template element*)

Il s'agit de trouver des caractéristiques d'un objet. Par exemple, pour un produit, trouver son nom, la société qui le fabrique, son prix, etc.

- La découverte d'une relation (*template relationship)*

Il s'agit de mettre à jour des relations entre les objets. On peut classer l'extraction d'informations sur les interactions dans cette catégorie.

- La description d'un évènement (*scenario template*)

Il s'agit de donner les caractéristiques d'un événement dans un texte : objets impliqués et modalités de réalisation. Par exemple : qui arrive à quel poste dans quelle entreprise. Dans le domaine de la génétique cela pourrait être : quelle protéine se fixe sur quel site promoteur, à quel stade du développement et quel est le type d'expérience qui a permis d'en apporter la preuve. Nous n'avons pas connaissance de système aussi élaboré dans le domaine de la génétique.

Nous voyons que l'extraction d'information se concentre sur la compréhension du texte, en évitant de faire référence à des connaissances extérieures. Par exemple, dans la tâche de reconnaissance des personnes, la question est plus de savoir qu'une portion de texte est un nom de personne (reconnaissance d'entité nommée) ou de savoir que plusieurs portions de textes renvoient en fait à un même individu (résolution d'anaphore), plutôt que d'identifier la personne avec une entrée d'un annuaire. Ainsi, la question est de comprendre ce qui est dit dans le texte et non de relier les éléments de compréhension acquis dans le texte à des connaissances acquises indépendamment.

Si le problème consiste à connecter, d'une part, des éléments de compréhension acquis dans le texte à, d'autre part, des connaissances du domaine acquises indépendamment, on se trouve face à d'autres exigences. On est placé dans le domaine de ce que nous appellerons la création de liens entre données factuelles et données bibliographiques. Cette thématique rejoint celles du couplage des bases de données, qu'elles soient bibliographiques ou factuelles.

IV. LE COUPLAGE DES BASES DE DONNÉES

Les informations utiles aux biologistes sont souvent disséminées dans de multiples bases de données, chacune de ces bases de données ayant ses propres buts (DISCALA *et alii*, 2000). Il serait illusoire de vouloir toutes les rassembler, de façon à présenter un point de vue unique sur la réalité. Apparaît néanmoins la nécessité de faire des liens entres les différentes informations qui s'y trouvent (KARP *et alii*, 1996).

A. LA CRÉATION DE LIENS ENTRE BASES DE DONNÉES

Il est vrai que chaque base de données s'emploie à créer des liens vers d'autres bases de données. Ainsi, il est très courant, dans les bases de données factuelles, de voir des références à des notices bibliographiques issues de *Medline*. Cependant, le travail de mise en relation est souvent fait manuellement. Il ne peut donc être exhaustif et rapide.

Des bases de données, créées spécifiquement pour compulser des données trouvées dans d'autres bases de données ont vu le jour. On peut citer ***Genecards***[8] qui est une encyclopédie sur les gènes humains (REBHAN *et alii*, 1998). Cette base de données rassemble sous une forme conviviale des informations sur les gènes, les protéines, les séquences et les pathologies. L'intérêt de cette base réside dans le fait que ces quatre types d'objets sont intimement liés. Plus précisément, la base est organisée autour des gènes qui sont présentés sous forme de 'carte', qui est un écran de synthèse sur toutes les informations rassemblées sur le gène. La base comporte 22400 cartes, ce qui correspond à 7000 gènes différents. Les informations sont issues des bases de données les plus complètes sur leurs sujets, entre autres *SWISS-PROT*, *OMIM*, *GENATLAS* et *GDB*.

OMIM[9] est un catalogue de gènes humains et de maladies associées. ***GENATLAS***[10] est une compilation d'information sur la cartographie du génome humain. ***GDB***[11] est une autre base de données sur la cartographie du génome humain.

DBGET/LinkDB est un autre projet d'acquisition et de gestion de données venant de bases de données biologiques hétérogènes (FU JIBUCHI *et alii*, 1998). C'est le système sur lequel s'appuie le programme ***GenomeNet*** dont fait partie l'encyclopédie sur les voies métaboliques *KEGG*.

Au niveau français, on peut citer la base de données prototype ***Virgil***[12] qui permet de gérer des liens entre deux bases de données, à savoir, ***GenBank***, qui est une base de séquences de gènes humains, et ***GDB*** qui est une base de données sur la cartographie des gènes humains (ACHARD *et alii*, 1998). Cette base de données permet de compter les liens entres les enregistrements des bases de données, de savoir s'ils sont réciproques ou unidirectionnels,

---

[8] REBHAN M, CHALIFA-CASPI V, PRILUSKY J, LANCET D. GeneCards: encyclopedia for genes, proteins and diseases. Weizmann Institute of Science, Bioinformatics Unit and Genome Center (Rehovot, Israel), 1997. Accessible sur http://thr.cit.nih.gov:8081/cards

[9] Accessible sur http://www3.ncbi.nlm.nih.gov/omim/

[10] Accessible sur http://www.citi2.fr/GENATLAS/

[11] Accessible sur http://gdbwww.gdb.org/

[12] Accessible sur http://www.infobiogen.fr/services/virgil/

etc. La base de données est accessible à travers ***CORBA***[13] qui est une interface qui permettant d'interroger plusieurs bases de données à la fois.

Pour faciliter les échanges d'informations entre les bases de données, les ontologies sont utilisées. Il s'agit de représentations formelles d'un domaine scientifique donné. Dans ces représentations, les objets et les relations entre ces objets sont décrits. Les ontologies peuvent être utilisées pour représenter les schémas d'organisation de l'information dans une base. L'ontologie devient la référence commune à plusieurs bases de données et elle permet l'échange d'information entre les différentes bases.

B. COUPLAGE DES BASES DE DONNÉES FACTUELLES AVEC DES BASES DE DONNÉES BIBLIOGRAPHIQUES

Dans une base de données, les informations peuvent être codées, soit dans un ***langage naturel***, par exemple l'anglais, soit codé dans un langage symbolique ou numérique. Dans le premier cas, nous parlerons de ***données textuelle***s, dans le second cas de ***données factuelles***. Nous parlerons de ***bases de données textuelles*** pour les bases de données qui renferment essentiellement du texte. Il s'agit principalement de bases de données bibliographiques telles que *Medline*. Les ***bases de données factuelles*** sont les bases de données qui contiennent peu de texte. Il s'agit par exemple de bases de données de séquences ou d'encyclopédies sur les gènes d'un organisme particulier comme *Flybase*.

Nous inscrivons notre travail dans le couplage de bases de données bibliographiques avec des bases de données factuelles et plus généralement dans la mise en relation de données textuelles et de données symboliques ou numériques. Il s'agit de mettre en relation un ensemble de textes avec une collection d'informations contenues dans une base de données. Les liens peuvent être considérés, soit comme un repérage de faits décrits dans des textes, soit comme le repérage d'un commentaire textuel sur des faits. Ainsi, il s'agit, soit de construire automatiquement une bibliographie, soit de repérer des faits décrits dans des textes.

La création de liens entre données textuelles et données factuelles touche à la fois à la recherche d'information et à l'extraction d'information :

- Quel sont les textes qui décrivent le mieux tel ou tel fait ?
- Quelles sont les meilleures représentations du contenu du texte ?

Dans notre cas, la base de données textuelles est *Medline*. Les bases de donnée factuelles que nous utilisons sont, d'une part *Flybase* et d'autre part une base de données sur les interactions génétiques que nous avons construite. Dans cette dernière, les interactions sont décrites par un couple de gènes et une indication sur le sens de l'interaction.

Le lien entre *Flybase* et *Medline* consiste à identifier un gène dans une phrase : tel gène apparaît sous telle forme dans telle phrase à tel endroit dans la phrase. Le lien entre la base de données sur les interactions que nous avons créée et *Medline* correspond à la reconnaissance d'une interaction dans une phrase : tel gène interagit avec tel gène (avec indication éventuelle du sens) d'après telle phrase.

Ainsi, chaque lien est de nature assez différente. Dans le cas de l'identification des gènes, la base de données sur les gènes est donnée d'avance et il s'agit donc davantage de recherche

[13] Voir http://www.corba.org/

d'information. Dans le cas de la reconnaissance des interactions, la base de données sur les interactions est créée au fur et à mesure et il s'agit donc davantage d'extraction d'informations.

Cependant, à l'avenir, nous aimerions travailler avec une liste d'interactions déjà constituée. Par exemple, il pourrait s'agir d'interactions hypothétiques obtenues par des expériences sur *puces à ADN*. La question serait alors d'avantage de trouver de la bibliographie permettant de valider les interactions plutôt que d'extraire de l'information.

Nous allons maintenant aborder l'état de l'art en ce qui concerne la question de la reconnaissance des gènes ou de leurs produits dans les textes. Dans la partie suivante nous traiterons de l'état de l'art en ce qui concerne l'extraction d'informations sur les interactions à partir de textes.

Chapitre 2 Études des travaux comparables

I. TRAVAUX CONCERNANT LA RECONNAISSANCE DE GÈNES OU DE LEURS PRODUITS DANS DES TEXTES

La reconnaissance des interactions génétiques ou moléculaires nécessite de savoir reconnaître les gènes dans les textes. Il s'agit aussi de savoir reconnaître le produit des gènes et notamment les protéines. Nous emploierons le terme de ***label*** pour tout terme relatif à un gène ou ses produits. Les travaux que nous présentons ici sont relatifs à la reconnaissance des *labels*. Ceux-ci s'intègrent pour la plupart dans des dispositifs plus larges d'analyse automatique de publications scientifiques. Les méthodes de reconnaissance des *labels* n'étant qu'un aspect secondaire des travaux présentés, elles ne sont pas souvent évaluées par leurs auteurs. Ces méthodes, inspirées de la tache de *reconnaissance d'entité nommée*, visent à repérer des portions de textes qui correspondent à des *labels*. Elles ne visent pas à identifier le gène, c'est à dire à mettre en relation le *label* avec le ou les gènes qui peuvent lui correspondre.

On distingue deux grands types de méthodes. Les premières utilisent des listes de *labels*. Les secondes essaient de reconnaître les occurrences de labels sans utiliser aucune connaissance sur la nomenclature. Dans cette section, nous discuterons aussi des travaux sur la création automatique de dictionnaires de gènes ou de protéine à partir de corpus.

A. TRAVAUX SUR LA RECONNAISSANCE DES GÈNES OU DE LEURS PRODUITS BASÉS SUR L'UTILISATION DE LISTES DE TERMES

La méthode la plus simple pour reconnaître l'utilisation d'un nom de gène dans un texte est l'utilisation d'un ***lexique***, autrement dit d'une liste non structurée de termes.

D'une façon générale, nous emploierons le terme de lexique pour désigner une simple liste. Dans le cas où la structuration des données est suffisante pour permettre de savoir que plusieurs termes désignent la même entité, nous emploierons le terme de ***dictionnaire***. Nous réserverons le terme de ***nomenclature*** dans le cas où les entités elles-mêmes sont structurées en classe. Il s'agit par exemple de protéines organisées en familles.

L'avantage d'un dictionnaire sur un lexique est qu'il rend possible l'identification précise d'un gène, alors que l'utilisation d'un lexique ne le permet pas.

Il existe de nombreux systèmes d'extraction d'informations sur les interactions génétiques ou moléculaires et plus généralement sur la fonction des gènes. Les auteurs prennent souvent le parti de ne travailler que sur un ensemble de gènes défini à l'avance. Ils renoncent de ce fait à des études globales sur le génome, mais cela leur permet de se concentrer sur les problèmes que pose la nomenclature des gènes. Dans ce cas, l'utilisateur a la possibilité d'intervenir sur les dictionnaires utilisés et de rajouter, par exemple, des synonymes qui manqueraient.

ANDRADE *et alii* dans leur travail d'extraction de mots clefs et de phrases clefs décrivant au mieux une famille de protéines, utilisent une simple liste de protéines (2001). La liste contient des *noms synonymes* mais les auteurs remarquent qu'ils ne sont pas tous répertoriés dans la liste qu'ils utilisent, et que cela oblige l'utilisateur à compléter manuellement la liste.

Cependant dans un autre travail auquel ANDRADE a participé, portant cette fois sur l'extraction d'informations sur les interactions entre protéines, les auteurs parviennent à s'abstraire de ce problème en travaillant sur un petit nombre de gènes (BLASCHKE *et alii*, 1999). Dans ce travail, les auteurs construisent des graphes de labels co-occurrents, c'est à dire des labels qui apparaissent dans les mêmes textes.

Le fait que le système soit utilisé sur un petit nombre de gènes permet de travailler avec une liste de synonymes incomplète. Dans ce système, *Medline* est interrogé avec ce petit nombre de protéines, que l'on sait être impliquées dans un même processus. La lecture des résumés obtenus permet de rajouter des synonymes à la liste de départ. Une nouvelle interrogation de *Medline* est alors faite avec le nouvel ensemble de noms de protéine.

Nous remarquons que le système proposé ne prend pas intégralement en compte le phénomène de synonymie, puisque dans le graphe ce n'est pas les protéines qui sont représentés mais les labels. Ainsi, il se peut que deux nœuds différents du graphe concernent en réalité la même protéine.

Le problème des homonymes (protéine ayant le même nom) est aussi remarqué par les auteurs. Là encore l'interrogation sur un petit nombre de protéines permet de contourner le problème. Les éventuels gènes homonymes sont, sauf cas exceptionnel, impliqués dans des phénomènes très différents du phénomène étudié. Ils seront donc cités dans d'autres résumés. Ainsi, il n'existera pas de résumé citant à la fois un gène d'intérêt et un gène homonyme. Le graphe de gènes co-occurrents, qui est le résultat final de l'analyse, sera donc exact.

PathBinder, qui est un système d'extraction d'informations sur les interactions, est un exemple de système qui se concentre sur une liste de gènes définie par avance (QI *et alii*, 2000). La recherche sur un ou plusieurs gènes donnés est élargie grâce à des listes de synonymes. Ces synonymes sont extraits de la nomenclature maintenue par le ***HUGO Gene Nomenclature Committee*** et par la base de données *OMIM*. Chaque synonyme est présenté à l'utilisateur pour vérification.

HUGO est une organisation internationale qui organise la coopération autour du séquençage et de la cartographie du génome humain. Elle possède un comité pour aider à la standardisation des noms de gènes. Ce comité rédige des recommandations et maintient une base de données sur la nomenclature des gènes.

Medminer est un système de recherche d'information sur la fonction des gènes et leurs relations à partir de résumés *Medline* (TANABE *et alii*, 1999). Il permet de sélectionner des résumés sur la base de la présence de certains mots clefs et de certains gènes ou couples de gènes. Ce système prend en compte la synonymie grâce aux informations extraites de *Genecards*. Les synonymes sont présentés à l'utilisateur pour validation. De ce fait, le système est adapté à la recherche sur un petit nombre de gènes.

A l'inverse des travaux présentés précédemment, ***PubGene*** travaille d'emblée sur un très grand nombre de gènes, ce qui lui permet de présenter des résultats basés sur des statistiques (JENSSEN *et alii*, 2001). *PubGene* est un système d'extraction d'informations sur les relations entre les gènes humains. Il travaille à partir de résumés issus de *Medline*. Il exploite la cooccurrence, c'est à dire le fait que plusieurs gènes soient cités dans le même texte. Il est donc important que les alias soient reconnus et correctement attribués aux gènes associés. Le système prend donc en charge la synonymie. Les informations sur la

nomenclature des gènes humains ont été obtenues par compilation de données provenant de différentes bases de données. Les bases de données utilisées sont : la base de donnée du ***HUGO Gene Nomenclature Committee***, ***GDB, GENATLAS*** et ***LocusLink***[14]. Cette dernière est une base de données sur la localisation chromosomique des gènes.

STEPHENS *et alii* proposent un système analogue (2001). Il s'agit d'extraction d'informations sur les relations qu'entretiennent les gènes. Le système utilise aussi un lexique défini avant toute expérience de nom de gènes ou de protéines.

Cependant la tâche d'identification des gènes dans les textes n'a pas été évaluée en tant que telle. C'est le résultat final, à savoir le réseau des gènes co-occurrents qui est évalué. Ce réseau est évalué du point de vue de sa pertinence pour le biologiste.

RINDFLECH *et alii*, dans leur travail en *recherche d'informations* sur les liaisons moléculaires entre macro-molécules, détectent les noms des objets en interactions grâce à leur fonction grammaticale dans la phrase et les identifient à des entrées de ***GenBank*** quand cela est possible (1999). Dans son travail sur l'extraction de relations entre médicaments, gènes et cellules, les noms de gènes sont reconnus comme tels grâce à l'utilisation d'un thésaurus (RINDFLESCH *et alii*, 2000). Ce thésaurus, ***l'UMLS Metathesaurus*** (HUMPHREYS *et alii*, 1998), est spécialisé dans le domaine médical. Il lui permet de reconnaître des objets de type cellule, médicament ou gène et de faire la distinction entre ces trois types d'objets. Pour les gènes, une liste de *noms synonymes* est adjointe à l'aide de *GeneCards*.

L'université de Tokyo développe un système d'extraction d'informations sur les interactions protéine-protéine (ONO *et alii*, 2001). Dans ce système, la reconnaissance des noms de protéines se fait par l'utilisation d'un dictionnaire sur les protéines. Ce dictionnaire a été créé semi-automatiquement par une analyse de la littérature sur laquelle nous reviendrons dans la partie réservée à l'étude de la bibliographie sur la création automatique de dictionnaire à partir de textes (YOSHIDA *et alii*, 2000).

En France, on peut noter le travail de TURNER *et alii* sur la création de liens entre *SwissProt* et *Medline* (2000). Dans ce travail, les résumés sont indexés avec des mots clefs extraits de SwissProt. Cette indexation permet de créer des liens d'un résumé vers des données factuelles contenues dans SwissProt. Les noms de protéines sont utilisés comme mots clefs pour indexer les résumés *Medline*. Le système est évalué du point de vue de la représentation documentaire. La question est de savoir si l'indexation des résumés est pertinente du point de vue d'une interrogation documentaire. Les auteurs n'ont pas évalué, en terme de rappel et de précision, leur technique de reconnaissance des protéines.

B. TRAVAUX SUR LA RECONNAISSANCE DES GÈNES OU DE LEURS PRODUITS N'UTILISANT PAS DE LEXIQUES

Pour éviter d'avoir à construire des dictionnaires ou des lexiques spécifiques au domaine, un certain nombre d'auteurs mettent au point des méthodes qui n'en nécessitent pas. Ces méthodes sont basées sur le fait que les noms de gènes ont une place spécifique dans la construction grammaticale de la phrase. Elles utilisent aussi les propriétés morphologiques des noms de gènes ou de protéines.

L'argument utilisé pour ne pas utiliser de dictionnaire est le suivant : les dictionnaires appropriés n'existent pas toujours. Une méthode générale doit donc pouvoir s'en passer.

[14] Accessible sur http://www.ncbi.nlm.nih.gov/LocusLink/

De plus, quand ces dictionnaires existent, ils ne sont pas à jour étant donné la quantité de nouveaux gènes et de nouvelles protéines découvertes chaque jour.

Ces arguments ne manquent pas de pertinence, tant il est vrai que les dictionnaires, quand ils existent, doivent être complétés et adaptés à la tâche de la reconnaissance ou de l'identification des gènes ou de leurs produits. Cependant, nous remarquons que les mêmes auteurs qui emploient ces arguments (FUKUDA *et alii*, 1998), élaborent aussi des programmes permettant de créer de tels dictionnaires automatiquement par l'analyse informatique des textes (YOSHIDA *et alii*, 1998). La non-disponibilité des dictionnaires spécialisés n'est donc pas un obstacle insurmontable. Nous verrons dans la partie réalisation, comment dans notre étude nous avons pu adapter le dictionnaire qui était à notre disposition.

Ces travaux s'inscrivent dans la tradition du traitement automatique des langues et plus précisément dans la tâche de *reconnaissance d'entités nommées*. Il s'agit de travaux sur *Medline*, sauf dans le cas du travail de PROUX (voir plus bas).

THOMAS *et alii*, dans leur travail sur les interactions entres protéines réalisent une analyse grammaticale des phrases (2000). Cela leur permet de détecter les syntagmes nominaux qui sont de bons candidats pour des noms de gènes. Ils utilisent des particularités morphologiques des noms de protéines telles que la présence de caractères spéciaux comme / - ( ) ou de chiffres qui sont souvent présents dans le nom des protéines auxquelles ils ont à faire. Cela leurs permet de se passer totalement de lexique sur les noms de gènes.

FUKUDA *et alii*, dans leur travail en *reconnaissance d'entités nommées* sur les protéines, utilisent les mêmes principes (1998). En particulier, ils utilisent le fait que les noms de protéines sont souvent en majuscules et comportent des caractères spéciaux et des chiffres. Les auteurs remarquent les difficultés introduites par la présence de *noms synonymes*. D'ailleurs, ils ne renoncent pas totalement à l'utilisation d'un dictionnaire des protéines même s'ils n'en utilisent pas dans cet article. Ils promettent des développements sur la construction automatique de dictionnaires de protéines. Le travail en question sera publié par YOSHIDA *et alii*. Nous y reviendrons à la section suivante qui est consacrée à la construction de dictionnaire par extraction d'informations dans des textes.

Tous ces auteurs travaillent sur des noms de protéines et non sur des noms de gènes. Or les noms de protéines ont des particularités morphologiques que n'ont pas les noms de gènes de la drosophile. Nous verrons que les noms de gènes chez la drosophile sont assez quelconques, c'est à dire que ce ne sont pas des noms de code. Ceci est moins vrai cependant pour les *symboles* qui sont des abréviations.

PROUX *et alii* ont effectué un travail sur la détection des noms de gènes, alors que les précédents auteurs ont travaillé sur les noms de protéines (1998). Cependant, ce travail est fait sur des textes issus de *Flybase*. ***Flybase*** est une base de données sur la drosophile sur laquelle nous reviendrons. Les textes en question, ont été écrits ou réécrits par les annotateurs de la base de données. Une terminologie stricte a été utilisée. Un seul nom est utilisé pour chaque gène. Plus précisément, il s'agit du *symbole* attribué par *Flybase* dans son dictionnaire des gènes. Le problème de la synonymie se trouve donc être artificiellement absent du corpus étudié. De plus, les *symboles*, qui sont des abréviations, ont des particularités morphologiques spécifiques, et ces particularités sont utilisées lors de la détection des labels. En outre, les *symboles* sont toujours composés d'un seul mot, ce qui

n'est pas le cas des *noms complets* et des *synonymes*. Ainsi ce travail n'est pas directement transposable aux textes issus de *Medline*.

PROUX *et alii* sont néanmoins confrontés à un problème intéressant qui est celui des *labels ambigus*. Ce sont des labels qui peuvent éventuellement désigner autre chose que des gènes. Ils font une distinction entre différents types de labels ambigus. Les labels ambigus hors du domaine (*out of scope*) sont des termes qui peuvent être caractérisés comme ambigus en général mais ne le sont pas dans le contexte des textes étudiés ici. Par exemple *gypsy*, qui signifie bohémien, n'est pas ambigu dans des textes de génétique sur la drosophile. Les labels ambigus dans le domaine (*in scope*) sont des termes qui peuvent prêter à confusion avec des termes du domaine. Il s'agit par exemple de *dorsal* qui est le nom d'un gène connu mais est aussi un terme anatomique. La dernière catégorie de label (*in conflict*) rassemble les noms de gènes qui prêtent à confusion avec un mot d'une autre catégorie grammaticale. Il s'agit par exemple de *is*, *a* , *by*, *red*, *can*. Ces termes, quand ils sont employés dans leurs sens de gènes, risquent de perturber l'analyse grammaticale de la phrase.

L'analyse grammaticale des phrases apporte néanmoins des informations pertinentes puisque les performances calculées par PROUX *et alii* sont intéressantes pour un système sans dictionnaire de gène. Le *taux de précision* atteint 91,4 % pour un *taux de rappel* de 94,4 %. A notre avis, ces techniques gagneraient à intégrer malgré tout un dictionnaire des gènes ; d'autant plus que la construction de tel dictionnaire à partir de corpus est possible. Les travaux décrits dans la section suivante le montrent.

RINDFLESCH *et alii* adoptent une stratégie dans laquelle les termes sont trouvés par analyse grammaticale de la phrase, puis éventuellement associés à des entrées d'une ressource terminologique (1999). Le travail porte sur l'extraction d'informations sur les affinités de liaisons moléculaires entre macro-molécules. Les termes recherchés sont tous les syntagmes nominaux qui peuvent éventuellement être sujets à une liaison moléculaire. Il peut donc s'agir d'une molécule, d'une partie d'une molécule, d'une cellule, d'une partie d'une cellule ou d'une structure génomique. Pour identifier ces entités, les auteurs utilisent des ressources terminologiques variées. Il s'agit en particulier du thésaurus biomédical ***UMLS Metathesaurus***, du dictionnaire biomédical ***SPECIALIST*** et de *Genbank*. Quand il n'y a pas de correspondance, le terme est laissé non interprété et le processus d'extraction d'informations suit son cours normalement. Il s'agit donc d'une démarche intermédiaire entre le 'tout lexique' et le 'sans lexique'.

C. TRAVAUX SUR LA CRÉATION AUTOMATIQUE OU ASSISTÉE DE DICTIONNAIRE DES GÈNES OU DE LEURS PRODUITS À PARTIR DE TEXTES

Les travaux que nous décrivons ici concernent la création de dictionnaire de gènes ou de protéines à partir de l'analyse de corpus de publications. Tous ces travaux proviennent de l'université de Tokyo.

Le premier travail concerne la création de base de connaissances à partir de publications scientifiques (OHTA *et alii*, 1997). Il a conduit à la réalisation d'un système qui combine recherche d'informations, extraction d'informations et construction de dictionnaires spécifiques. Le logiciel s'appelle *IFBP* pour *Information Finding from Biological Paper*. Le dictionnaire spécialisé utilisé par le système contient entre autres des noms de protéines. Ce dictionnaire sert à analyser les textes pour la recherche et l'extraction d'informations. Un intéressant mécanisme de rétroaction est mis en place. Le résultat de la recherche d'informations permise par le dictionnaire sert à améliorer le dictionnaire lui-même. Le

système est utilisé pour construire un thésaurus spécialisé. La méthode appliquée est la classification hiérarchique des termes. Dans cette classification, les termes sont regroupés s'ils partagent les mêmes contextes dans la littérature.

Signalons que l'extraction de terminologie se heurte à un problème classique en traitement automatique des langues : la résolution de la ***collocation***. Il s'agit de reconnaître quand plusieurs mots contigus forment une unité de sens. Le système proposé est capable de les résoudre par des méthodes statistiques (YAMAMOTO *et alii*, 1996). Cela permet d'extraire des noms de protéine qui sont composés de plusieurs termes.

Dans ce travail, les auteurs soulignent le rôle central que jouent les dictionnaires spécialisés pour l'analyse des textes de biologie.

YOSHIDA *et alii* proposent un système d'aide à la création de dictionnaire de noms de protéines (1998). Il s'agit de la création d'un dictionnaire d'acronymes.

Un des problèmes dans la manipulation de lexiques est que l'on ne sait pas quand deux termes distincts désignent la même chose. Souvent la relation entre les deux termes est un lien d'abréviation. L'objectif du travail de YOSHIDA *et alii* est de repérer ces relations d'abréviation quand elles sont explicitement indiquées dans les textes. Ce travail est très utile, car nous avons pu constater que ce type de relations est effectivement indiqué de façon tout à fait explicite. Concrètement, les auteurs donnent le nom abrégé immédiatement après le *nom complet* et entre parenthèse. L'auteur cite en exemple *Thyrotrophin-releasing hormone (TRH)*.

Le logiciel présenté par YOSHIDA *et alii* détecte la présence de ces explicitations d'abréviation, qu'il nomme *parenthetical paraphrase*. Le logiciel repère la présence des parenthèses et la relation d'abréviation (YOSHIDA *et alii*, 2000). Cette relation d'abréviation peut consister à prendre les premières lettres de chaque mot comme dans l'exemple ci-dessus. Il y a cependant d'autres cas de figures qui sont pris en charge par le système. Cette méthode a été utilisée pour créer un dictionnaire sur les protéines de la levure *Saccharomyces* et de la bactérie *Escherichia coli*. Ces dictionnaires sont mis en œuvre pour extraire des informations sur les interactions entre protéines (ONO *et alii*, 2001).

Dans un échantillon de 112 résumés que nous avons annoté avec précision, nous avons pu constater que 62 résumés sont concernés par ce phénomène, soit plus de la moitié. L'explicitation d'abréviation utilisée par YOSHIDA *et alii* est donc fréquente, ce qui rend son travail extrêmement intéressant.

Nous verrons dans la partie réalisation (section Partie 2) comment nous utilisons, nous aussi, le phénomène. Il s'agira pour nous, soit de valider des interprétations possibles du texte, soit de compléter automatiquement le dictionnaire dont nous disposons.

Concernant la construction de nomenclature de gènes rigoureuse, nous signalons le travail de LICCIULLI (1999) et de CATALANO (2000) sur les séquences génétiques. Ce travail consiste à mettre de l'ordre dans la terminologie utilisée pour décrire les séquences dans les bases de données de séquences. Les auteurs ont travaillé sur la base de données de séquences nucléique de l'***EMBL***[15] (STOESSER, 2001). Dans ce type de base de données, les séquences sont associées à des noms de gènes et à des noms de protéines mais la

[15] Accessible sur http://www.ebi.ac.uk/embl/

nomenclature utilisée n'est pas extrêmement rigoureuse. Cela est dû au fait que le vocabulaire n'est pas contrôlé. Les auteurs sont libres dans la description des séquences qu'ils soumettent. Chacun va utiliser son propre vocabulaire pour énumérer gènes et protéines associés à la séquence. Le résultat est qu'un gène ou une protéine peut avoir été désigné par différents noms. Inversement un même nom peut désigner des gènes ou des protéines qui n'ont rien à voir. Les auteurs distinguent deux causes dans l'inconsistance de la terminologie. La première est biologique. Une même protéine pourra être désignée de différentes façons selon le contexte biologique. La seconde est sémantique. Il existe plusieurs orthographes possibles pour le *nom complet* d'une protéine et plusieurs façons d'abréger ce nom.

Ce manque de rigueur dans la description des séquences par des mots clefs rend l'information disponible plus difficilement exploitable. L'auteur propose des solutions à ce problème. Il procède par classification des mots clefs. Cette classification va rassembler des mots clefs associés à des séquences identiques ou impliquées dans des fonctions similaires. La navigation dans cette classification doit permettre à l'utilisateur de mieux connaître la terminologie utilisée pour désigner les gènes qui l'intéressent. Ce type de recherche est donc exploitable pour construire des dictionnaires de gènes, mais cette construction n'est pas automatique, elle est simplement assistée par l'informatique.

D. CONCLUSION SUR LES TRAVAUX CONCERNANT LA RECHERCHE DE GÈNES OU DE LEURS PRODUITS DANS DES TEXTES

La majeure partie des travaux sur la reconnaissance de labels porte sur les protéines. Une originalité de notre travail consiste à s'intéresser aux noms des gènes et de leurs produits.

Les études faites sur la reconnaissance des gènes ou de leurs produits permettent d'isoler un certain nombre de problèmes. Premièrement, la présence de nombreux *noms synonymes* vient compliquer l'identification d'un gène précis dans une liste préalablement définie. Deuxièmement, l'existence de gènes homonymes peut constituer une source d'erreurs. Troisièmement, la présence de *labels ambigus* pose le problème d'une reconnaissance contextuelle.

Rares sont les systèmes qui prennent intégralement en charge le phénomène de synonymie. La plupart des systèmes se contentent de poser une marque à chaque occurrence d'un label. Ils ne se préoccupent pas d'associer le label à un gène unique en prenant en compte le fait que plusieurs labels puissent renvoyer à un même gène.

Les systèmes proposés utilisent des listes de *noms synonymes* extraites de base de données mais nous n'avons pas trouvé d'études quantitatives sur les carences de ces listes.

Les conséquences de l'homonymie dans la nomenclature et de l'ambiguïté de certains labels ne semblent pas non plus avoir été quantitativement évaluées.

Ainsi, il manque une étude sur la possibilité d'utiliser une base de données pour identifier les gènes dans les textes. Cette base de données devrait contenir les divers *noms synonymes* des gènes et de leurs produits.

Cette étude doit permettre de montrer qu'il est possible d'identifier des gènes dans des textes sans faire appel à des techniques linguistiques infiniment plus sophistiquées que l'utilisation d'un dictionnaire.

Elle doit aussi permettre de quantifier les conséquences des problèmes cités ci-dessus, à savoir :

- Incomplétudes des listes de synonymes disponibles
- Existence de gènes homonymes
- Existence de labels ambigus

Elle pourra aussi permettre de quantifier des phénomènes tels que :

- Les non-utilisations de la nomenclature officielle (fréquence de l'utilisation de synonymes à la place des noms canoniques)
- L'utilisation des *noms abrégés* à la place des *noms complets*
- L'explicitation dans le texte des relations entre *noms complets* et abréviations (voir *explicitation d'abréviation*)

Par ailleurs, les travaux en extraction d'informations, et plus précisément en *reconnaissance d'entités nommées* et en *reconnaissance d'acronymes* permettent d'envisager la création automatique de dictionnaires de gènes. Ces développements permettraient alors de résoudre le principal problème de la méthode que nous proposons qui est celui de la disponibilité de dictionnaires complets et à jour.

II. TRAVAUX SUR L'ANALYSE INFORMATIQUE DES TEXTES ET LES INTERACTIONS GÉNÉTIQUES OU MOLÉCULAIRES

Le second volet de notre travail consiste à extraire des informations sur les interactions génétiques et moléculaires. Cette tâche relève de l'*extraction d'informations* telle que nous l'avons définie précédemment. La plupart des travaux appartiennent au domaine de la compréhension du langage naturel qui est une branche de l'intelligence artificielle qui elle-même est une discipline de l'informatique. Il s'agit de rechercher des motifs spécifiques de la description d'une interaction. Il existe différentes approches que nous présenterons successivement.

Nous traiterons des méthodes basées sur la cooccurrence des gènes. La cooccurrence est un facteur qui est utilisé dans tous les travaux mais qui n'est pas toujours mis en valeur. Il semble que les auteurs considèrent la cooccurrence comme un critère allant de soi. De fait, il est naturel de penser qu'un texte qui décrit une interaction entre deux gènes, les citent tous les deux au moins une fois. La cooccurrence apparaît comme une condition nécessaire mais non suffisante. La plupart des auteurs vont se concentrer sur ce que leur analyse des textes peut apporter comme information supplémentaire. Cependant, certains travaux utilisent la cooccurrence comme source d'informations principales. Il s'agit, dans ce cas, de détecter des corrélations statistiques entre l'apparition d'un gène –dans un texte –et l'apparition d'un autre gène. Ces corrélations expriment des relations fonctionnelles évidentes ou révèlent des relations cachées. Ces études permettent donc l'exploration de connaissances bien établies tout en proposant une aide à la découverte.

Nous traiterons aussi des méthodes basées sur la recherche de mots clefs ou de phrases clefs. Ces travaux sont fondés sur le repérage des segments de textes pertinents pour une question donnée.

D'une part, les mots clefs sont utilisés par la plupart des méthodes pour sélectionner les textes qui vont être traités. D'autre part, les travaux en compréhension du langage naturel n'utilisent pour la plupart d'entre eux qu'un tout petit nombre de mots pivots, souvent des verbes, pour repérer les interactions. Ainsi, les méthodes classiques utilisent d'une certaine

façon des mots clefs même si elles se concentrent sur ce qu'elles peuvent apporter en plus au processus d'extraction d'informations. Nous réserverons donc une partie aux méthodes qui traitent de l'utilisation de mots clefs pour permettre l'accès à l'information pertinente sur les interactions.

A. MÉTHODES D'ANALYSE INFORMATIQUE DES TEXTES SUR LES INTERACTIONS GÉNÉTIQUES ET MOLÉCULAIRES BASÉES SUR LA RECHERCHE DE MOTS CLEFS ET DE PHRASES CLEFS

Nous présentons dans cette section des travaux appartenant au domaine de la recherche d'informations. Leur but est de faciliter l'accès au texte en repérant des points saillants ou en classant les textes selon des thèmes. Ces travaux ont été menés sur *Medline*. Les auteurs s'intéressent aux interactions entre les gènes ou leurs produits et plus généralement à la fonction des gènes. La méthode utilisée est celle de la recherche de mots clefs et de phrases clefs.

AbXtract[16] est un système de recherche d'informations sur la fonction des protéines (ANDRADE *et alii*, 1998). Il permet de sélectionner dans la littérature les phrases les plus informatives sur la fonction d'une famille de protéines donnée. Ces phrases, appelées phrase clefs, sont repérées selon des critères statistiques (ANDRADE *et alii*, 1997). Il s'agit de savoir si une phrase donnée contient des mots, appelé mots clefs, plus spécifiquement associés à la famille d'intérêt qu'à d'autres familles. Les textes sont présentés à l'utilisateur avec un code de couleurs qui lui permet de visualiser les éléments du texte les plus significatifs. Nous présentons un exemple ci-dessous.

Exemple 1 Détection de phrases clefs et de mots clefs par le logiciel *AbXtract*.
Les mots clefs sont en gras, les phrases clefs sont soulignées. La première phrase correspond au titre de l'article. Le résumé présenté est celui dont le numéro est 96362658. Le formatage est calculé par *AbsXtract* avec la requête *cap2*, qui est le nom d'une protéine chez la levure.

***Mutational** analysis of **capping** protein function in **Saccharomyces cerevisiae**. To investigate physiologic functions and structural correlates for **actin capping** protein (CP), we analyzed site-directed **mutations** in **CAP1** and **CAP2**, which encode the alpha and beta **subunits** of CP in **Saccharomyces cerevisiae**. **Mutations** in four **different** regions caused a **loss** of CP function in **vivo** despite the presence of **mutant** protein in the cells. **Mutations** in three regions caused a complete **loss** of all aspects of function, including the **actin distribution**, viability with sac6, and **localization** of CP to **actin cortical** patches. **Mutation** of the fourth region led to partial **loss** of only one function-formation of **actin cables**. Some **mutations** retained function and exhibited the complete **wild**-type **phenotype**, and some **mutations** led to a complete **loss** of protein and therefore **loss** of function. The simplest hypothesis that can explain these **results** is that a single biochemical property is necessary for all in **vivo** functions. This biochemical property is most likely binding to **actin filaments**, **because** the nonfunctional **mutant** CPs no longer co-localize with **actin filaments** in **vivo** and **because** direct binding of CP to **actin filaments** has been well established by studies with purified proteins in **vitro**. More complex hypotheses, involving the existence of additional biochemical properties important for function, cannot be excluded by this analysis.*

Nous avons remplacé le codage en couleur par un codage en noir et blanc. On voit que le titre a été repéré comme phrase clef, plus quatre phrases au milieu du résumé.

Les mots clefs et les phrases clefs ont été déterminés de la manière suivante. Pour une famille de protéines donnée, on considère l'ensemble des résumés qui traitent de cette

[16] Accessible sur http://www.pdg.cnb.uam.es/blaschke/cgi-bin/abx

famille. Pour un terme donné, on définit F comme la proportion des résumés qui utilisent le terme. La moyenne et l'écart type de la variable F sont calculés sur les familles à terme constant. On définit le score du terme dans la famille par la formule :

$$\tilde{\chi} = \frac{F - \overline{F}}{\sigma}$$

Dans le cas où le terme ne serait utilisé que dans une famille, il ne serait pas possible de calculer l'écart type. Dans ce cas, on prend un score égal à dix fois F (ANDRADE *et alii*, 2001). Les mots clefs sont, par définition, les mots dont le score est supérieur à un certain seuil. Le score de la phrase est obtenu en faisant la moyenne des scores des mots qui la composent. Les phrases clefs sont les phrases dont le score est supérieur à un certain seuil.

Cette technique, qui permet de trouver des mots clefs et des phrases clefs caractéristiques d'un ensemble de protéines, a aussi été utilisée pour interpréter les données d'expressions des gènes. BLASCHKE *et alii* (2000) proposent le système ***GEISHA***[17]. Ces données d'expression sont le résultat d'expériences sur *puces à ADN*. La comparaison des données d'expressions permet d'identifier des groupes ayant des profils semblables. Ces données de classification doivent être interprétées. L'utilisation de la littérature est un moyen d'obtenir des mots clefs caractéristiques de chaque groupe. Des méthodes de statistiques textuelles similaires à celles précédemment exposées sont utilisées. L'interface proposée à l'utilisateur permet de lier les mots clefs aux résumés et donc au contexte d'utilisation de ces mots clefs. Les mots clefs sont aussi reliés entre eux de façon à pouvoir naviguer entre les mots clefs pour explorer l'échantillon étudié.

La relation entre profils d'expression comparable et similarité fonctionnelle a été étudiée par JUAN CARLOS OLIVEROS *et alii.* (2000). Les auteurs montrent que des gènes ayant mêmes profils d'expression auront des contextes textuels semblables dans *Medline*. Cette proximité dans les textes signe d'après les auteurs une similarité fonctionnelle.

La méthode mise en œuvre dans *AbXtract* a été adaptée à la mise à jour automatique des informations contenues dans les bases de données. Il s'agit d'extraire de la littérature des mots clefs pour décrire des données présentes dans la base de données *OMIM* (ANDRADE, 2000). L'utilisation de la littérature permet de garantir la fraîcheur des données.

BLASCHKE *et alii* ont aussi utilisé *AbXtract* pour mettre au point un système d'extraction d'informations sur les protéines (1999). Le logiciel a servi à établir une liste de verbes pertinents. Nous y reviendrons dans la suite.

TANABE *et alii* proposent ***Medminer***[18], un système de recherche d'informations sur les gènes et leurs implications dans des pathologies (1999). Ce système permet de rechercher et de classer des résumés issus de *Medline*. Ce logiciel est donc une aide à la lecture. Il permet de mettre en avant des textes pertinents pour une question particulière. Cette question est définie par avance. Il s'agit de la fonction des gènes, de leurs interactions et de leurs implications dans des pathologies. La sélection et le classement des résumés se font selon des mots clefs. Un petit nombre de thèmes de classement est proposé. Pour chacun de ces thèmes, une petite liste de mots clefs est dressée. Les textes qui utilisent ces mots clefs sont sélectionnés et regroupés dans les thèmes correspondants. Les thèmes et les mots clefs

[17] Accessible sur http://www.pdg.cnb.uam.es/blaschke/cgi-bin/geisha

[18] Accessible sur http://discover.nci.nih.gov/textmining/main.html

associés sont définis avant toute expérience, contrairement aux méthodes présentées précédemment. Ce sont donc toujours les mêmes, alors que dans *AbXtract* les mots clefs sont différents pour chaque famille de protéines. Un des thèmes pris en charge correspond à l'interaction moléculaire et les mots clefs associés sont : *bind**, *cataly**, *clear** et *transcri**, où l'astérisque représente une troncature. Un autre thème correspond à l'inhibition et les mots clefs associés sont : *Downregulat**, *block**, *deplet**, *deficien**, *decreas**, *inhibit**, *reduc** et *absen**. Comme on peut le constater, les mots clefs sont souvent des verbes. L'utilisateur a la possibilité d'interroger le système avec des noms de gènes ou de protéines. Le système est adapté à l'étude des relations que peuvent entretenir deux gènes entre eux grâce à la possibilité qui est offerte d'effectuer une interrogation avec un couple de labels. Dans ce cas, une phrase va être considérée comme pertinente si elle contient un des deux labels et si un mot clef est utilisé. Un résumé sera considéré comme pertinent s'il contient une phrase pertinente et si les deux labels sont cités. Le résultat d'une interrogation est consultable en hypertexte.

B. MÉTHODES BASÉES SUR DES ÉTUDES STATISTIQUES D'APPARITION DE MOTS CLEFS POUR EXTRAIRE DES INFORMATIONS SUR LES INTERACTIONS GÉNÉTIQUES OU MOLÉCULAIRES

Un autre courant de recherche utilise la statistique textuelle pour obtenir des informations synthétiques sur la fonction des gènes. Les statistiques sont effectuées sur des mots du texte ou sur des descripteurs.

SHATKAY *et alii* proposent un système pour caractériser des ensembles de gènes par des mots clefs (2000). Les auteurs font appel à des modèles statistiques qui décrivent la fréquence des mots dans les documents. Les documents sont représentés par les mots qu'ils utilisent. La question qui est posée est celle de trouver les termes qui représentent le mieux un ensemble de documents donnés. Le système est utilisé pour interpréter des données d'expression issues de *puces à ADN*.

MASYS *et alii* proposent un système analogue d'interprétation de données d'expressions (2001). Il s'agit aussi d'interpréter les résultats d'expériences issues de *puces à ADN*. Les groupes de gènes, ayant des profils similaires, sont caractérisés grâce à des données issues de *Medline*. Les informations extraites de *Medline* sont les termes *MeSH* qui servent à l'indexation des résumés. Ainsi les profils d'expressions sont associés à des termes *MeSH*. Les auteurs utilisent en particulier la classification hiérarchique des enzymes que propose le *MeSH*. Nous avons là un exemple très intéressant de liens établis entre des données d'expériences et des données de type encyclopédique. Ces liens sont établis par l'intermédiaire des publications. Il s'agit de composer deux liens. Le premier lien va des données d'expérience vers les données de publications. Il s'agit de trouver les publications qui donnent des informations sur les gènes dont on a des données d'expressions. Le deuxième lien va des publications vers des données de classification. Il s'agit d'une description des documents faite par *Medline*. Le résultat est une description très riche des données d'expériences.

Biobibliometrics[19] est un système d'extraction d'informations sur les fonctions des gènes basé sur des statistiques de cooccurrences des labels (STAPLEY *et alii*, 2000). Le système est basé sur le fait que les gènes n'apparaissent pas « au hasard » dans les textes. Il y a des corrélations. Le traitement statistique vise à découvrir ces corrélations. Des couples de

[19] Accessible sur http://www.bmm.icnet.uk/~stapleyb/biobib/

gènes en relation sont ainsi mis en évidence. Les auteurs affirment que ces corrélations sont liées à des similitudes fonctionnelles. Le système permet aussi de révéler des relations qui ne sont pas évidentes au premier abord. Il permet donc la découverte. Le résultat est présenté sous forme de graphes. Le système est interrogeable à partir de mots clefs. Pour résumer, le système permet de connaître les gènes impliqués dans un phénomène donné et leurs relations de cooccurrence.

STEPHENS *et alii* (2001) proposent aussi un système basé sur des statistiques de cooccurrence pour extraire des informations sur les relations qu'entretiennent les gènes entre eux. Les relations en question ne sont pas définies avec précision. Il peut s'agir d'interaction, de participation à des processus communs ou simplement de partage de caractéristiques communes. La démarche est donc clairement celle de la découverte. Un graphe est construit automatiquement pour visualiser le résultat d'une requête. Dans ce graphe, les nœuds représentent des labels et les branches représentent des relations de cooccurrence. La longueur d'une branche est d'autant plus petite que les labels ont tendance à être présents dans les mêmes résumés. Cependant, l'importance d'une cooccurrence dans le calcul va dépendre de l'importance des gènes concernés dans la représentation du document à l'intérieur de l'espace des documents. L'interprétation des graphes n'est donc pas évidente. Le système proposé permet en plus de caractériser la relation entre les gènes. La caractérisation de la relation qu'entretiennent deux gènes co-occurrents est réalisée comme suit. Une liste de descripteurs possibles est dressée avant toute expérience. Il s'agit de mots clefs que l'on est susceptible de trouver dans les résumés. Parmi ces descripteurs, est choisi, pour caractériser la relation celui qui est le plus statistiquement significatif de l'ensemble de résumés qui co-citent les gènes. Les expériences sont menées sur un petit groupe de gènes que l'on sait être en relation. Le dispositif est d'un maniement assez délicat puisque le nombre de résumés doit être sensiblement le même pour chaque label.

PubGene[20]est un système analogue qui exploite la cooccurrence pour faire des statistiques (JENSSEN *et alii*, 2001). Ce logiciel est dédié à l'étude des relations que peuvent entretenir les gènes humains entre eux. Ce système exploite plus de 10 millions de notices bibliographiques issues de *Medline*. Les cooccurrences sont recherchées dans le titre ou dans le résumé. Le système permet aussi d'étudier des relations plus larges. Il s'agit de trouver des gènes présents dans des articles co-cités. La co-citation, c'est à dire, pour deux articles, le fait d'être cité en référence bibliographique dans un troisième article, révèle une relation entre les deux articles. Les gènes décrits dans des articles en relation sont en relation eux-mêmes. C'est cette relation entre les gènes qui est étudiée. Pour faire ce travail les auteurs ont utilisé le ***Science Citation Index***. Dans cette base de données la bibliographie de chaque article est incluse dans la notice (QUONIAM, 1996).

Le premier but de *PubGene* est de visualiser des graphes de gènes en relation. Cependant, les auteurs proposent d'autres types d'utilisation, à savoir :

[20] Accessible sur http://www.pubgene.org/

- parcourir la littérature associée à un gène donné de façon plus ou moins large,
- rechercher la littérature relative à un groupe de gènes,
- rechercher des termes associés à un gène donné,
- trouver les noms officiels d'un gène donné,
- rechercher les termes *MeSH* associées à un groupe de gène,
- interpréter des données d'expressions.

C. MÉTHODES BASÉES SUR LA COOCCURRENCE POUR EXTRAIRE DES INFORMATIONS SUR LES INTERACTIONS GÉNÉTIQUES OU MOLÉCULAIRES

RINDFLESCH *et alii* propose un système d'extraction d'informations sur les relations qu'entretiennent gènes, médicaments et cellules (2000). Il s'agit de trouver des relations du type : dans les cellules de type C, l'expression du gène G est inhibée (ou activée) par le médicament M, ou du type, les cellules du type C acquièrent une résistance (ou une sensibilité) au médicament M quand le gène G s'exprime. Ce type d'informations est utile dans l'étude du cancer. Le système est basé sur la reconnaissance dans une même phrase d'un gène, d'un type cellulaire et d'un médicament. Il s'agit donc d'un système basé sur la cooccurrence, même si la relation en question est tripartite. Le système proposé ne permet pas de déterminer le type de relation entre les trois entités mais les auteurs projettent d'y arriver à l'avenir.

PILLET *et alii* proposent aussi un système basé sur la cooccurrence (1998). Il s'agit d'extraire des informations sur les interactions génétiques ou moléculaires à partir de commentaires bibliographiques contenus dans *Flybase*. Le système est basé à la fois sur la cooccurrence et sur la présence d'un vocabulaire spécifique dans la phrase. Nous avons utilisé ce travail et nous donnerons plus d'information quand nous décrirons nos réalisations.

D. MÉTHODES BASÉES SUR DES MOTIFS TEXTUELS POUR EXTRAIRE DES INFORMATIONS SUR LES INTERACTIONS GÉNÉTIQUES OU MOLÉCULAIRES

Pour détecter des interactions, la cooccurrence de deux labels n'est pas un facteur suffisant. La cooccurrence peut avoir lieu pour bien d'autres raisons que la description d'une interaction. Des éléments supplémentaires d'informations doivent être adjoints pour décider si une interaction est décrite ou non. De plus, dans le cas où il y aurait cooccurrence, et sauf dans le cas où il n'y aurait que deux labels, il resterait à déterminer entre quels labels les interactions ont lieu. La question du sens, de A vers B ou de B vers A, doit aussi être posée. La question de savoir quel est le type de l'interaction, activation ou inhibition par exemple, reste elle aussi ouverte.

Ainsi, il apparaît nécessaire de faire une analyse plus approfondie de la phrase. Il est important de repérer des verbes tel que *activate* ou *inhibit* et de déterminer sujets et compléments d'objets associés. Il s'agit donc de faire appel au traitement automatique des langues. Les travaux qui suivent font appels à cette technique.

BLASCHKE *et alii* (1999) proposent un système d'extraction d'informations sur les interactions entres protéines. Ce système s'appelle ***suiseki***[21]. Il permet de travailler sur un ensemble prédéfini de protéines impliquées dans un même processus. Les résumés issus de *Medline* qui correspondent à ce processus sont analysés. Il s'agit de reconnaître des motifs du type : Protéine A – Action – Protéine B, où Protéine A et Protéine B sont des noms de

[21] Des exemples d'application sont accessible sur http://www.pdg.cnb.uam.es/suiseki/index.html

protéines et Action est un verbe appartenant à une liste prédéfinie. Les auteurs recherchent aussi des phrases du type « la protéine A est un membre de la famille B » et ils utilisent d'autres motifs textuels pour cela. Le verbe utilisé pour l'action permet aussi de déterminer quel est le type de l'interaction. Le système est testé dans des cas pratiques, ce qui permet aux auteurs de démontrer sa pertinence. Cependant les auteurs n'ont pas évalué la précision et le rappel.

SEKIMIZU *et alii* proposent un système d'extraction d'information sur les interactions entre les gènes ou leurs produits (1998). Les interactions sont détectées grâce à la présence de verbes tels que *activate*, *bind*, *interact*, *regulate* et *inhibit*. Ce travail s'inscrit dans le projet GENIA d'acquisition de connaissances à partir de publications sur le génome (1999). Les auteurs évaluent la précision de leur résultat mais ne sont pas en mesure d'évaluer le rappel.

L'université de Tokyo développe un système d'extraction d'informations sur les interactions entre protéines (ONO *et alii*, 1999). Les motifs utilisés sont du type *Protéine A – interact with – Protéine B*. Le système a été testé sur la levure *Saccharomyces cerevisiae* et sur la bactérie *Escherichia coli*. Les taux de rappel atteignent respectivement 86,8% et 82,5% (ONO *et alii*, 2001). Cependant les performances du système sont évaluées sur les seules phrases qui contiennent les motifs recherchés. Ainsi le rappel calculé ne prend pas en compte toutes les interactions qui ont été manquées à cause d'un motif trop spécifique. Les taux de précision atteignent respectivement 94,3 % et 93,5 %.

La même équipe de recherche présente aussi un travail sur les voies biologiques (HISHIGAKI *et alii*, 1999). Il s'agit d'extraire des informations sur les relations entre les protéines et les voies biologiques dans lesquelles la protéine est impliquée. Les auteurs proposent d'exploiter les données sur les interactions pour obtenir des informations sur la fonction des protéines et les voies biologiques.

D'autres équipes de recherche travaillent sur l'extraction d'informations sur les voies biologiques. Le système ***PIES*** (***Protein Interaction Extraction System***) associe des fonctions de recherche et d'extraction d'informations sur les interactions entre protéines, de manipulation des informations sur les interactions et sur les voies biologiques et de visualisation des réseaux d'interactions (NG *et alii*, 1999). Dans ce système la recherche d'informations se fait par repérage de certains verbes. Ces verbes sont choisis avant toute expérience. Ils sont regroupés en classes selon des thèmes. Le thème *Inhibition* regroupe les verbes *inhibit*, *suppress* et *negativly regulate*. Le thème *Activation* regroupe les verbes *activate*, *transactivate*, *induce*, *upregulate* et *positively regulate*. Les phrases contenant un de ces verbes sont sélectionnées. Des motifs sont recherchés dans ces phrases de façon à extraire des informations sur le réseau d'interactions décrit. Le stockage et la manipulation des informations sur les interactions et les voies biologiques est réalisé dans une base de données. Le système permet aussi d'intégrer des informations sur les interactions issues de calcul sur les séquences (WONG *et alii*, 2001). Il s'agit de repérer des évènements de fusion entre gènes au cours de l'évolution. *PIES* se présente donc comme un système complet d'acquisition, de manipulation de visualisation d'informations sur les voies biologiques.

Les systèmes d'extraction d'informations sur les interactions peuvent être des adaptations d'un système plus généraliste. Le système d'extraction d'informations ***Highlight*** est une adaptation du système ***FASTUS*** développé par le ***SRI*** (THOMAS *et alii*, 2000). Les informations extraites concernent les interactions entre protéines. Les verbes utilisés sont *interract*, *associate* et *bind*. Un test effectué sur un échantillon de 90 résumés fait ressortir des taux de rappel et de précision de respectivement 29 % et 69 %.

Chapitre 3 Notre apport et celui du consortium Cerise

I. HISTORIQUE DES TRAVAUX DANS LE CONSORTIUM CERISE

Nous présenterons dans cette partie le consortium dans lequel nous avons travaillé et les expériences qui y ont été mené avant ou parallèlement à notre travail.

A. PRÉSENTATION DU PROGRAMME DE RECHERCHE DU CONSORTIUM CERISE

Notre travail s'inscrit dans le programme du ***Consortium d'Etude des Réseaux d'Interactions des Systèmes Eucaryotes (Cerise)***. Nous allons en présenter le schéma général avant de voir plus en détail certains résultats obtenus dont nous avons besoin pour définir le cadre de notre travail.

Cerise a pour but l'étude des interactions génétiques et moléculaires. Il fait partie du programme génome du CNRS. Cinq équipes de recherche venant de disciplines différentes y prennent part. La plupart de ces équipes font partie du CNRS et travaillent sur des thématiques de génétique. Deux laboratoires font exception car leurs thèmes de recherche ne sont pas traités au CNRS. Il s'agit de l'*INRIA Rhône-Alpes*[22], spécialisé en intelligence artificielle, et du *CRRM* qui est spécialisé en science de l'information et dont je fais partie. La compagnie *Xerox* a participé aux opérations en collaborant avec l'*INRIA Rhône-Alpes*.

La direction du consortium est assurée par BERNARD JACQ qui est chargé de recherche au *LGPD*[23].

Pour mener son étude sur les interactions génétiques et moléculaires, le consortium s'est donné trois axes de travail :

- Saisie d'informations sur les interactions génétiques et moléculaires
- Représentation de ces informations à partir de base de connaissance par objet
- Exploitation de ces informations : analyse, comparaison et simulation de fonctionnement.

Voyons maintenant brièvement chacun de ces axes de recherche.

1. La saisie des informations sur les interactions génétiques et moléculaires

Cette tâche, à laquelle nous avons participé, vise à mettre à disposition du consortium et plus spécialement de la communauté des chercheurs en génétique, un nombre important d'informations tirées d'articles scientifiques sur les interactions génétiques et moléculaires. Deux techniques différentes ont été utilisées.

La première, mise en œuvre avec l'aide du CRRM est basée sur l'utilisation des statistiques textuelles. Ce travail a été initié en 1996 lors du DEA de PILLET et a été complété par une thèse (2000). Cette technique a servi de point de départ à notre travail et nous reprendrons un certain nombre de résultats et en particulier la mise en évidence d'un vocabulaire

[22] Institut national de la recherche en informatique et automatique.

[23] Laboratoire de génétique et développement de Marseille.

spécifique aux phrases décrivant une interaction génétique ou moléculaire ainsi que le critère de la présence simultanée de deux noms de gènes dans une même phrase.

La seconde technique utilisée, basée sur la linguistique informatique a été mise en œuvre par l'***INRIA Rhône Alpes*** et la Compagnie ***Xerox***. Elle a donné lieu à une thèse (PROUX, 2001)

Ces deux techniques d'extraction d'information ont été développées indépendamment. Cependant elles sont complémentaires au sens où la seconde pourrait s'enchaîner à la première.

2. La représentation des connaissances sur les interactions génétiques et moléculaires

Elle vise à la construction de modèles informatiques aptes à représenter la réalité biologique des interactions génétiques et moléculaires. Elle est mise en œuvre grâce à des techniques de bases de connaissances par objet. Cet axe de recherche a été mené par l'*INRIA Rhône-Alpes* en collaboration avec la Compagnie *Xerox*. Pour plus de précisions, je renvoie le lecteur à la thèse de DENYS PROUX (2001).

3. Analyse, comparaison et simulation de fonctionnement des réseaux régulateurs

La capitalisation des connaissances sur les interactions génétiques permet d'envisager une compréhension fine de la façon dont les gènes coopèrent à la réalisation d'une même fonction. Il est en effet possible de construire des modèles mathématiques de certain réseau de gènes en interactions. Ces modèles permettent de simuler le fonctionnement de ces réseaux dans leur dynamique. Il est par exemple possible de prédire l'existence de cycles limites ou d'états stationnaires. De tels modèles mathématiques sont généralement très simplificateurs. Ainsi dans ces modèles, un gène est soit exprimé, soit non exprimé à un instant donné alors que dans la réalité les gènes peuvent avoir des effets différents selon leurs degré d'expression. De même, dans ces modèles, les interactions ne font intervenir que deux gènes alors que dans la réalité il peut arriver que trois molécules, voir d'avantage, interagissent physiquement. Pour finir, ces modèles considèrent les groupes de gènes étudiés comme isolés de toute influence extérieure. Cependant de tels modèles, malgré leur simplicité, permettent par exemple de décrire le fonctionnement complexe d'une dizaine de gènes impliqués dans la formation des fleurs chez la plante *Arabidopsis thaliana* (MENDOZA, 1999).

D'un organisme à l'autre, la comparaison des réseaux de gènes en interaction permet de donner des informations sur l'évolution fonctionnelle des gènes. C'est un des buts que se donne le consortium avec notamment la réalisation de la base HOFLYPROT[24] qui est une base de données des gènes *homologues*.

B. CHOIX MÉTHODOLOGIQUE INITIÉ PAR PILLET

1. Choix de la base de données Flybase

Pour tester les méthodes d'extraction d'information, PILLET a choisi d'utiliser la base de données *Flybase*. Voyons les raisons de son choix.

[24] Voir http://pbil.univ-lyon1.fr/databases/hoflyprot.html

a. Présentation de la base de données Flybase

Flybase est une base de données spécialisée sur la drosophile. Elle répertorie entre autres des références bibliographiques. Pour chaque référence bibliographique, le texte intégral de la publication est lu par un annotateur de *Flybase*. Les informations, qui y sont contenues, sont relatées et classées dans des catégories pré établies. Cette structure permet d'avoir accès à l'information de différentes manières et pas uniquement par référence bibliographique. Un accès par gène permet notamment de connaître la synthèse de tout ce qui a été publié sur un gène donné. Un champ a particulièrement reçu l'attention de PILLET, il s'agit du champ *Phenotypic information*. Les informations sur les interactions génétiques ou moléculaires sont fréquentes dans ce champ. Ce champ se présente comme un résumé de quelques lignes. Les informations se présentent donc sous une forme semblable à celles qui sont issues de *Medline*. Cependant, l'utilisation de ce champ présente de nombreux avantages comme nous allons le voir dans la section suivante.

b. Avantages de la base de données Flybase

La base de données *Flybase* fait autorité et est très complète en ce qui concerne les interactions génétiques ou moléculaires chez la drosophile. On peut donc s'attendre à y trouver l'essentiel des informations.

Dans les textes issus de *Flybase*, l'expression de l'information est beaucoup plus compacte et homogène. En effet, les rédacteurs des textes écrivent des petits résumés de quelques lignes sur les articles qu'ils ont lus. A l'opposé, les résumés issus de *Medline* font environ une demi-page. La concision de *Flybase* est très avantageuse pour mettre en œuvre les techniques d'extraction d'informations : plus petite est la proportion de texte non pertinent, moins on a de chance de considérer à tort une partie du texte comme pertinente.

Le fait que les textes issus de *Flybase* soient écrits par quelques personnes seulement contre plusieurs centaines d'auteurs différents pour les textes issus de *Medline*, permet de garantir une certaine homogénéité dans l'expression. Ceci est très précieux pour déceler par des méthodes statistiques des motifs spécifiques d'expression d'interactions génétiques ou moléculaires.

Par ailleurs, certaines caractéristiques de *Flybase* permettent d'éviter des erreurs d'analyse qui demeurent possibles dans les textes issus de *Medline*. Par exemple, sur *Medline* il peut y avoir des informations sur des gènes d'autres espèces et ce, d'autant plus que les noms de gènes sont parfois les mêmes d'une espèce à l'autre. Ceci est évité dans *Flybase* puisqu'elle se consacre presque exclusivement à la drosophile. Par ailleurs, l'identification des gènes dans les textes issus de *Medline* pose des difficultés spécifiques alors que dans *Flybase* les annotateurs s'imposent une terminologie stricte et parfaitement décrite dans *Flybase* elle-même. C'est un point important, et nous reviendrons sur les difficultés de l'identification des gènes dans les textes issue de *Medline* à la section Partie 2.

2. Choix d'une méthode d'analyse basée sur la présence conjointe de noms de gènes et d'un vocabulaire spécifique dans une même phrase

La méthode qui a été adoptée est très intuitive, car elle est basée sur un constat simple et pragmatique sur les textes à analyser : les phrases qui décrivent une interaction génétique ou moléculaire se caractérisent par un vocabulaire spécifique et la présence de plusieurs occurrences de gènes.

Voyons-le sur l'exemple suivant :

Exemple 2 Phrase extraite de Flybase qui décrit une interaction
Il est dit dans cette phrase qu'un certain type d'analyse (l'analyse par double mutant) permet d'établir que le gène *sdt* agit en aval du gène *crb* et est activé par celui-ci.

Double mutant analysis suggests that sdt acts downstream of and is activated by crb.

Nous voyons par cet exemple que l'analyse à faire sur la phrase pour en comprendre le sens est assez poussée. Elle nécessite des connaissances approfondies du domaine. Il est implicitement question d'une voie de régulation génétique à laquelle les deux gènes participent. Des informations sont données sur la place et le rôle de chacun des deux gènes :

- Le sens du signal est précisé : il s'agit d'un signal de *crb* vers *sdt*.
- Le signe du signal est précisé : l'expression de *sdt* a une action positive sur l'expression de *crb*.

Une analyse à base d'intelligence artificielle serait donc difficile à mettre en œuvre. En revanche, nous remarquons que le simple fait de repérer des noms de gènes et des syntagmes tel que *act downstream* ou *is activated by* serait suffisante pour extraire l'information dont nous avons besoin. L'exemple suivant l'illustre.

Exemple 3 Traits caractéristiques servant à l'analyse
La présence simultanée dans la même phrase de deux noms de gènes et d'expressions aussi spécifiques que *act downstream of* et *is activated by* permet de conclure assez raisonnablement qu'une interaction est décrite et qu'elle met en jeu les deux gènes cités.

Double mutant analysis suggests that ***sdt*** *acts downstream of and is activated by* ***crb****.*

L'analyse des textes va consister à repérer à la fois un vocabulaire spécifique et des noms de gènes. Une interaction sera détectée quand une même phrase utilisera un vocabulaire spécifique et comptera au moins deux occurrences de noms de gènes, qui formeront alors les partenaires de l'interaction. Nous symboliserons ce principe par l'équation 1.

Équation 1 Le principe de l'analyse
Une interaction est décrite en faisant référence aux partenaires de celle-ci et par l'utilisation d'un vocabulaire spécifique, et vice-versa.

Interaction = Partenaires + vocabulaire spécifique

Voyons maintenant la méthode plus en détail, et notamment la méthode de détection du vocabulaire spécifique.

C. LA MÉTHODE DES IVI

Comme cela est illustré par l'équation ci-dessus, pour mener à bien notre analyse, il faut être capable de dire pour une phrase donnée et au vu de son vocabulaire, si elle est susceptible ou non de décrire une interaction. Voyons tout d'abord comment nous proposons d'isoler un vocabulaire spécifique, puis comment analyser globalement le vocabulaire d'une phrase pour décider de chances de la phrase de décrire une interaction.

1. Identifier le vocabulaire spécifique de l'interaction

Nous avons choisi d'aborder le sens des textes par l'analyse de leur seul vocabulaire. C'est à dire que nous ne voulons pas prendre en compte l'ordre des mots, ni même la présence simultanée de plusieurs mots. Ainsi, nous pensons que certains mots possèdent à eux seuls, en dehors de tout contexte ou de toute combinaison avec d'autres mots, un pouvoir de discriminer entre les phrases qui décrivent une interaction et celles qui n'en décrivent pas.

Downstream en est un bon exemple puisque l'on imagine mal que ce mot puisse servir à autre chose qu'à caractériser le place relative d'un gène par rapport à un autre dans une voie de signalisation. Dés lors, toute phrase qui utilise ce terme a de grande chance de décrire une interaction. Il se trouve qu'effectivement, dans le corpus étudié par PILLET, toutes les phrases qui utilisent *downstream* décrivent bien une interaction. Le tableau 1 donne les phrases qui utilisent le terme *downstream*.

Tableau 1 Notion de terme spécifique
Le tableau suivant donne toutes les phrases de *Flybase* qui utilisent le terme *downstream*. Toutes décrivent une interaction. Le terme est donc très spécifique.

Homozygous females and females of the genotype Df(1)HC244,Sxl[M1]/ovo are sterile with ovaries devoid of germ cells: ovo must act downstream of Sxl or in a different pathway
Females of the genotype Df(1)HC244,Sxl[M1]/fl(1)302 are sterile: fl(1)302 acts downstream of Sxl or in a different pathway
phl acts downstream of tor
csw functions downstream of tor
Jra is required downstream of the sev signalling pathway for development in the eye
Sor1 acts downstream of phl in the DER pathway
Genetic analysis suggests that pnt is a downstream effector of Ras85D
Kr acts downstream of ct in the Malpighian tubule regulatory pathway
srp acts downstream of hkb to promote morphogenesis and differentiation of anterior and posterior midgut
srp acts as a homeotic gene downstream of the terminal gap gene hkb to promote morphogenesis and differentiation of anterior and posterior midgut
Reduction in intensity of twi expression at gastrulation correlates well with degree of dorsalization of embryos, suggesting effect of dl mediated through its downstream target genes
The exd gene product acts with the selector homeodomain proteins, including Ubx, as a DNA binding transcription factor, thereby altering their regulation of downstream target genes
Double mutant analysis suggests that sdt acts downstream of and is activated by crb
Ems is a downstream gene that is transcriptionally regulated by Abd-B gene products
In-vivo ac is a direct downstream target of h regulation

Le terme *act* est lui aussi assez spécifique bien que les phrases qui utilisent ce terme ne décrivent pas toutes une interaction : 10 phrases sur un total de 60 ne décrivent pas d'interaction. Les proportions sont encore un peu moins favorables pour le terme *activated* puisque 14 phrases sur un total de 64 ne décrivent pas d'interaction.

Un tableau permettant de classer entre eux les termes associés à l'énoncé d'une interaction à été établi. Le tableau 2 en donne un extrait. Les calculs ont été faits, non pas sur les termes eux-même, mais sur les ***lemmes*** associés. La lemmatisation consiste à ramener les formes fléchies, à savoir les formes conjuguées et les pluriels, à des formes standardisées, à savoir l'infinitif ou le singulier.

Tableau 2 Vocabulaire spécifique d'une interaction
Les lemmes les plus spécifiques d'une interaction sont listés ici. La colonne fréquence donne le nombre de phrase qui utilisent au moins une fois le lemme. La colonne portion donne le nombre de phrase qui utilisent le terme et décrivent une interaction. Dans la colonne proportion on trouve le rapport entre les deux chiffres précédents.

Lemme	Fréquence	Portion	Proportion
synergistic	9	9	100%
positive	9	9	100%
downstream	15	15	100%
cardinal	6	6	100%
prefer	5	5	100%
autoregulate	11	11	100%
amnioserosa	7	7	100%
modulate	12	12	100%
epistasic	5	5	100%
derepress	7	7	100%
initiate	12	11	92%
class	30	27	90%
negative	19	17	89%
exert	9	8	89%
interact	149	131	88%
alter	24	21	88%
multiple	8	7	88%
ontogenetic	8	7	88%
sequence-motif	7	6	86%
pc-group	7	6	86%
zygote	44	37	84%
downregulate	6	5	83%
pre-mrna	6	5	83%
supply	6	5	83%
r7	6	5	83%
act	60	50	83%
trans-act	6	5	83%
transductor	6	5	83%

On voit apparaître le lemme *downstream* en haut du tableau et le lemme *act* en bas du tableau. En revanche, le dernier lemme de l'exemple *Activated* n'apparaît pas dans l'extrait présenté ici.

La proportion des textes qui décrivent une interaction parmi ceux qui utilisent un terme donné paraît être un bon indicateur de la spécificité du lemme. On nommera cette quantité ***spécificité***. La définition en est donnée en page 45.

Définition 1 Spécificité d'un lemme
N_{lemme} est le nombre de phrases qui utilisent le lemme, et n_{lemme} est le nombre de phrases décrivant une interaction qui utilisent le lemme.

$$Spécificité_{lemme} = n_{lemme} / N_{lemme}$$

2. Sélectionner les textes qui décrivent une interaction

L'idée la plus simple consiste à sélectionner un certain nombre de termes parmi les plus spécifiques et à exiger pour retenir une phrase que l'un au moins de ces termes soit présent. Cette technique ne prend malheureusement pas en compte le fait éventuel que plusieurs termes spécifiques peuvent être présents et elle s'est révélée avoir de faibles performances. Ainsi, la présence d'un terme, est à lui seul un indice insuffisant pour établir un diagnostic sur une phrase. Des méthodes, exigeant la présence simultanée de plusieurs termes particuliers, ont été testées par PILLET sous l'appellation *méthode des requêtes bi-termes ou pluri-termes*. Elles ont donné de moins bons résultats en terme de performances que la méthode dite des *IVI* qui va être détaillée ci-après.

Pour prendre en compte le concours que peut apporter chaque terme au diagnostic que nous cherchons à établir, nous pouvons imaginer faire la moyenne des spécificités pour chaque terme présent dans la phrase à analyser. On obtient alors pour la phrase un indice global que PILLET a nommé *IVI* pour Index de Vraisemblance d'Interaction. L'*IVI* est donc calculé selon la formule ci-dessous.

Définition 2 *IVI* d'une phrase
N_{phrase} est le nombre d'occurrences de termes spécifiques que compte la phrase.

$$IVI_{phrase} = \sum_{terme\ dans\ la\ phrase} Spécificité_{terme} \Big/ N_{phrase}$$

Les phrases sont ensuite comparées entre elles et les premières (par ordre l'*IVI*) sont sélectionnées. Autrement dit, on se donne un seuil et on sélectionne les phrases qui ont un *IVI* supérieur à ce seuil. Le choix de ce seuil est arbitraire. Cependant il est clair que les dernières sélectionnées, ayant un *IVI* moins favorable, ont plus de chance d'être sélectionnées à tort. Ainsi, plus on sélectionne de phrases, plus mauvaise est la qualité de la sélection.

Le principe de l'extraction d'information peut alors se résumer par l'équation ci-dessous où *s* est un seuil fixé au départ.

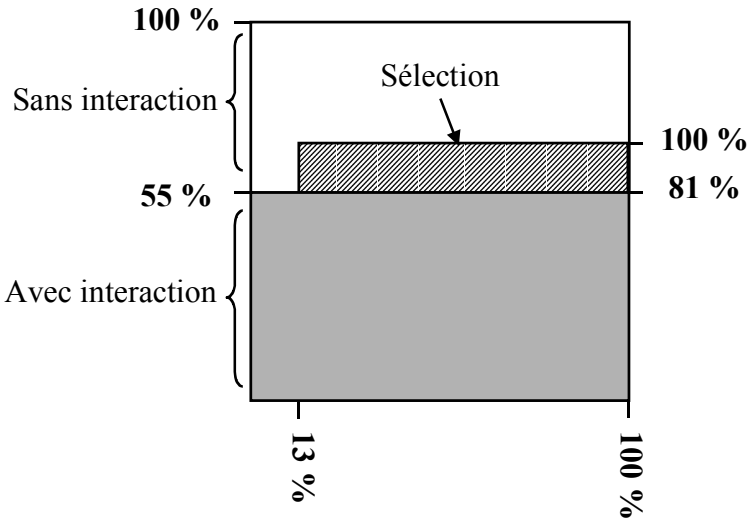
Équation 2 Principe de l'analyse par utilisation des *IVI*
Une interaction est décrite quand les partenaires sont cités et que l'*IVI* est supérieur à un certain seuil

$$Interaction = partenaires + IVI > s$$

3. Performance de la méthode des IVI

La méthode mise en œuvre par PILLET permet effectivement d'enrichir le corpus en phrases qui décrivent une interaction. En effet, avant application du critère de l'*IVI*, le taux de phrases qui décrivent une interaction est de 55%. Il passe à 81% après application du critère de l'*IVI* –le seuil choisi étant zéro. Cependant, cet enrichissement se fait moyennant une perte d'informations : 13 % des phrases qui décrivent une interaction sont écartées par erreur. Ceci est illustré par la figure 1.

Figure 1 Résultat de la méthode des *IVI*
Le grand carré représente l'ensemble des phrases avant la sélection. Le petit rectangle hachuré représente l'ensemble des phrases après la sélection. La zones grisée représente les phrases qui décrivent une interaction.



D. LES VARIANTES DE LA MÉTHODE DES IVI

PILLET a proposé plusieurs formules pour calculer l'*IVI*. Nous avons fait le choix d'une méthode parmi plusieurs et nous avons aussi fait quelques simplifications par rapport à ce qui a été proposé initialement. Nous allons ici justifier ces choix.

1. Variante dans le calcul de la spécificité

La définition de la spécificité utilisée par PILLET est différente de celle que nous proposons. Nous allons montrer que la définition que nous proposons est équivalente après une simplification que nous justifierons.

Nous avons défini la spécificité à la section C.1 et notamment par la définition 1. PILLET utilise la définition ci-dessous.

Définition 3 Définition de la spécificité utilisée par PILLET
N_{terme} désigne le nombre de phrases utilisant le terme ; n_{terme} désigne le nombre de phrases utilisant le terme qui décrivent une interaction ; n'_{terme} désigne le nombre de phrases utilisant le terme ne décrivant pas d'interaction.

$$Spécificité_{terme} = (n_{terme} - n'_{terme}) / N_{terme}$$

PILLET a classé les phrases en trois catégories : les phrases qui décrivent une interaction, celles qui n'en décrivent pas et enfin celles pour lesquelles la situation n'est pas suffisamment claire pour pouvoir trancher. Nous avons décidé de ne faire que deux catégories en plaçant les phrases de la catégorie des *indécises* dans la catégorie des *non*.

Le principal intérêt de cette modification consiste à faciliter la comparaison de notre travail avec d'autres travaux. En effet, dans la plupart des travaux sur l'extraction d'information, on ne distingue que deux catégories de texte –les bons et les mauvais –et non trois catégories.

D'autre part, cette modification permet de substituer la définition 3 par la définition 1 qui est à notre sens plus simple. En effet, ces deux formules sont équivalentes comme nous allons le montrer dans ce qui suit.

La présence de deux catégories seulement permet d'écrire $N = n + n'$. D'où, après calcul $S' = 2\,S - 1$ où S' désigne la spécificité définie par PILLET, et S la spécificité que nous proposons. On en déduit $IVI' = 2\,IVI - 1$ où IVI' désigne l'index de vraisemblance d'interaction calculé grâce aux spécificité S'. Les deux versions de l'*IVI* sont identiques à la composition par une fonction affine près. Ainsi, quand IVI' prend ses valeurs entre -1 et 1, IVI varie entre 0 et 1. Imposer une condition du type $IVI > s$ revient exactement à imposer une condition du type $IVI' > s'$ où s' est un seuil choisi en fonction du seuil s. Nous pouvons donc dire que les deux statistiques IVI et IVI' sont équivalentes. Nous utiliserons la définition 3 dans les calculs de l'*IVI* car elle a l'avantage de conduire à une quantité qui change de signe, ce qui est plus lisible dans les graphiques.

2. Calcul de l'IVI par la somme des spécificités

Une variante pour calculer l'*IVI* consiste à faire la somme à la place de la moyenne des spécificités, comme explicité ci-dessous.

Définition 4 Une variante pour la définition de l'IVI

$$IVI_{phrase} = \sum_{terme\ dans\ la\ phrase} Spécificité_{terme}$$

Cette formule a donné des résultats moins bons, comme illustré dans le tableau 3, où l'on voit qu'à précision égale, la méthode de la somme ne permet pas un aussi bon rappel.

Tableau 3 Prise en compte de la spécificité de chaque terme : somme ou moyenne
Le tableau donne la performance de la méthode d'extraction d'information. La performance est évaluée en terme de précision et de rappel. Deux méthodes existent pour prendre en compte la spécificité de chaque terme du texte pour obtenir chiffre global pour le texte considéré. La première consiste à faire la moyenne des spécificités, la seconde consiste à faire la somme.

Méthode	Précision	Rappel
Moyenne	81%	87%
Somme	81%	62%

Par ailleurs, la méthode de la somme a l'inconvénient de fournir un *IVI* qui n'est pas borné. En effet, la spécificité est comprise entre 0 et 1 ; quand on fait la moyenne, la quantité reste comprise entre ces deux bornes. En revanche si l'on fait la somme, on peut avoir des nombres beaucoup plus grands. Cela trahit le fait que l'*IVI* calculé par la méthode de la somme est une statistique, qui n'est pas indépendante de la taille du texte. En effet, si le texte est long, il comptera en moyenne plus de termes spécifiques que si le texte est court. Dans ce cas, l'*IVI* risque d'être plus grand. Cette non-indépendance de l'*IVI* par rapport à la taille du texte n'est pas pertinente car il n'y a pas lieu de penser que les grandes phrases ont plus de chance de décrire une interaction. Nous pensons que c'est la raison pour laquelle l'*IVI* calculé par la méthode de la somme donne de moins bons résultats.

3. Calcul de l'IVI par l'analyse factorielle

Une analyse factorielle des correspondances a été proposée par PILLET pour représenter l'ensemble des termes placés dans un espace à trois dimensions, où chaque coordonnée représente le nombre de phrases qui ont été jugées comme décrivant une interaction, n'en décrivant pas, ou n'en décrivant pas clairement. Le facteur principal donne alors la spécificité du terme. L'*IVI* est ensuite calculé selon la formule de la somme.

Cette méthode a donné des résultats très proches de celle que nous avons adoptée et décrite précédemment. Le tableau 4 permet une comparaison des performances :

Tableau 4 Calcul de la spécificité : proportion ou analyse factorielle
Le tableau donne la performance de la méthode d'extraction d'information selon la méthode de calcul de la spécificité utilisé. Les deux méthodes d'extraction d'information donnent des résultats comparables.

Méthode	Précision	Rappel
Proportion	81%	87%
Analyse factorielle	84%	85%

On constate que le calcul par analyse factorielle donne de meilleurs résultats en terme de précision mais de moins bons en terme de rappel. Nous pouvons donc dire que les méthodes se valent et nous avons choisi la méthode la plus intuitive de calcul de la spécificité : la proportion de phrases qui traitent d'interaction.

II. RÉFLEXIONS SUR LA MÉTHODE D'ANALYSE QUE NOUS PROPOSONS

A. CHOIX DU CORPUS D'ANALYSE

1. Choix de Medline

Medline est une base de données bibliographiques. Cela signifie qu'elle est conçue pour aider les chercheurs à se tenir informés des avancés scientifiques dans leur discipline. *Medline* recense des articles scientifiques de façon à donner accès à ces documents, le chercheur étant invité à commander le document dont il a lu un résumé dans la base de données. L'intérêt de ce type de base de données bibliographique réside pour nous dans le fait que les auteurs donnent souvent l'essentiel de leurs résultats dans le résumé. Les interactions décrites dans les articles eux-même ont donc toutes les chances de se trouver aussi dans le résumé. Par ailleurs, les articles sont souvent d'accès payant et il n'existe pas de base de données gratuite d'articles en texte intégral, tout du moins de taille comparable à *Medline.*

La base de données *Medline* à l'avantage d'être accessible gratuitement et d'être très complète. Elle compte en effet plus de neuf millions de résumés. Elle traite d'aspects très divers de la médecine et de la biologie. Cependant, même sur un aspect précis comme celui de la génétique de la drosophile, elle peut rivaliser avec les meilleures bases de données spécialisées. Nous en voulons pour preuve qu'elle est largement utilisée par les chercheurs s'intéressant à la génétique de la drosophile.

Son principal avantage pour notre travail consiste dans le fait qu'elle traite de plusieurs *organismes modèles.* Ainsi, la drosophile est traitée, mais aussi la souris et l'homme qui sont deux organismes modèles très intéressants pour la recherche sur les interactions génétiques ou moléculaires. Ainsi, les méthodes que nous mettons au point sur la drosophile pourront

être adaptées facilement à l'analyse de données sur d'autres organismes modèles. Cela n'aurait pas été le cas s'il avait fallu changer de base de données.

2. Choix de l'échantillon d'analyse

Nous nous intéressons au texte de *Medline* qui traite de la génétique de la drosophile en général. Cependant, il a fallu faire le choix d'un échantillon d'analyse. Nous expliquons ce choix dans cette section.

Notre échantillon est constitué des résumés issus de *Medline*, qui sont cités dans au moins une phrase du corpus de PILLET. En effet, les textes étudiés par PILLET sont tirés de la lecture de publications de résultats qui sont pour la plupart référencés par *Medline*. Ainsi, il existe un lien naturel entre ces phrases et les résumés de *Medline*.

Ce choix rend la comparaison avec le travail de PILLET plus facile. En effet, le corpus étudié par PILLET est naturellement riche en interaction, car il est obtenu par une présélection des phrases sur le critère de la présence de deux occurrences de gènes dans chaque phrase. Ce critère assez exigeant permet d'avoir dès le départ une bonne proportion de textes qui relatent des interactions. Il a d'ailleurs été appliqué pour cette caractéristique. Cette bonne proportion de texte qui relate des interactions est tout naturellement conservée dans le lien qui relie les textes étudiés par PILLET et ceux que nous avons inclus dans notre échantillon d'analyse.

La richesse comparable en énoncés d'interaction est utile pour la comparaison des performances car cette richesse affecte les taux de précision à rappel égal. Par exemple, pour un rappel de 100%, la précision n'est autre que le taux de phrases relatant une interaction.

Plus prosaïquement, ce lien entre le corpus d'étude de PILLET et le nôtre permet de voir si les interactions décrites par les opérateurs de *Flybase* sont ou non présentes dans les résumés associés. Il est en effet possible que certaines d'entre elles ne soient présentes que dans le texte intégral des publications. Inversement, il est intéressant de savoir s'il y a des interactions présentes dans les résumés issus de *Medline* qui ne sont pas reprises dans l'échantillon d'étude de PILLET. Dans ce cas cela signifierait que les annotateurs de *Flybase* auraient oublié de noter certaines informations présentes dans les résumés. Une autre solution serait que cette absence d'information dans le corpus de PILLET soit due à la méthode de constitution du son corpus. Je pense notamment au critère de présence simultanée de deux occurrences de gènes.

3. Utiliser les données issues de Flybase pour analyser les textes de Medline

Il peut sembler surprenant de vouloir utiliser des données issues de *Flybase* pour analyser des données issues de *Medline*. Il y a plusieurs justifications à cela.

Tout d'abord, comme nous l'avons évoqué à la section I.B.1.b, les données issues de *Flybase* sont plus homogènes que les données issues de *Medline*. Elles sont donc plus intéressantes pour obtenir le vocabulaire spécifique des interactions à partir de méthodes statistiques.

Ensuite, et c'est le principal, en utilisant les données issues du travail de PILLET, nous n'avons pas le problème de la distinction entre données d'apprentissage et données de test. Les résultats que nous obtenons peuvent être considérés comme des données de test alors que PILLET avait ce problème de l'absence de données de test. Nous évitons, par

l'utilisation des données statistiques obtenues sur un autre corpus, d'avoir à constituer des résultats réservés à l'apprentissage.

B. DISCUSSIONS SUR LES MOYENS ET LES BUTS

1. La présence de deux noms de gènes est un indice fort

La méthode que nous adoptons pour la reconnaissance des interactions est basée, d'une part sur la présence de noms de gènes dans les textes, et d'autre part sur la présence d'un vocabulaire spécifique. Ces deux facteurs sont importants mais on évalue mal l'importance relative de chacun. En effet, l'analyse faite par PILLET étudie l'importance du facteur vocabulaire sur un échantillon dans lequel une condition forte est déjà posée sur la présence de gène : il s'agit d'exiger que deux occurrences de gène exactement soient présentes dans une même phrase. Ce critère a été posé pour une raison pragmatique. Comment savoir en effet de quelles interactions il est question dans une phrase qui possède le vocabulaire spécifique s'il y a plus de deux acteurs en présence ou s'il n'y en a qu'un seul et que l'autre est sous-entendu ?

Le critère sur les gènes étant posé au préalable, on évalue mal son influence sur les résultats. Nous pensons que le critère appliqué par PILLET sur les noms de gène est une condition très forte et qu'elle est la cause principale de la réussite du système d'extraction d'information. Nous remarquons en effet que dans le corpus sélectionné par PILLET sur le critère des noms de gènes, 55% des phrases décrivent une interaction. Ainsi, cet ensemble de phrases est déjà très riche en interactions. On ne connaît pas ce que serait ce taux sans le critère, car le critère a été appliqué dés le départ, mais on peut penser qu'il serait très faible.

Ainsi, nous pensons que le critère sur la présence de gènes est primordial et donc que nous devons concentrer notre travail sur la réalisation d'un système performant d'identification des gènes dans les textes. A la fin de notre exposé, dans la partie résultat, section Partie 2Chapitre 3II.C, nous fournirons la preuve du fait que le critère sur la présence de gènes est principalement responsable de la réussite de la méthode. Cela nous est permis par une analyse humaine des textes qui se fait sans sélection au préalable des textes qui contiennent un nombre déterminé d'occurrence de gènes.

2. Utilisation des phrases qui comportent plus de deux occurrences de gènes

Nous proposons en effet de généraliser la méthode proposée par PILLET aux phrases qui comptent plus de deux occurrences de gènes. Cela revient à considérer tous les couples de gènes en présence dans la phrase comme candidats éventuels à une interaction.

L'avantage consiste, bien évidemment, à disposer de plus de matériel de départ : il y a assez peu de phrases qui comportent exactement deux occurrences de gène, et beaucoup plus qui en comptent d'avantage. L'inconvénient attendu est de générer davantage de faux positifs. Nous verrons exactement ce qu'il en est dans la partie résultat.

L'annotation des textes a été faite en ce sens : toutes les phrases ont été annotées, indépendamment du nombre d'occurrence de gènes. Cela nous permet d'évaluer, par la même occasion, l'importance du critère du nombre d'occurrences de gènes pour détecter les interactions.

3. Reconnaissance des interactions et non des phrases qui décrivent des interactions

PILLET a conçu un système qui permet la reconnaissance des phrases qui décrivent une interaction. Nous proposons un système qui permet la reconnaissance des interactions elles-mêmes. Le système que nous proposons va donc plus loin : il donne en plus des phrases qui traitent d'interactions, des informations sur les interactions dont il s'agit.

Il est vrai que dans le cas du corpus analysé par PILLET la différence entre les deux systèmes est faible. En effet, dans ce corpus chaque phrase comporte exactement deux occurrences de gènes, de sorte qu'il n'y a qu'une seule interaction qui puisse y être associée automatiquement.

Remarquons cependant que, même dans ce cas, la différence existe, puisque qu'il se peut que l'interaction réelle ne soit pas celle attendue. En effet, on ne peut pas exclure que l'interaction ne fasse intervenir qu'un des deux partenaires cités. Le cas peut se présenter quand il y a rétroaction ou quand un troisième partenaire n'est pas cité par son nom.

Dans notre corpus de textes annotés, la différence est importante puisqu'il n'y a plus correspondance entre une phrase et une interaction potentielle : une même phrase peut être la source de plusieurs reconnaissances automatiques d'interactions.

Partie 2

Réalisation et résultats

Chapitre 1 Analyse des problèmes posées

Nous analysons dans cette partie le problème posée par l'identification des gènes dans les textes et par l'extraction d'information sur les interactions. Nous expliquerons les méthodes que nous avons mises en œuvre pour les résoudre dans la partie 0.

I. INVENTAIRE DES DIFFICULTÉS À RÉSOUDRE POUR RÉALISER UN PROGRAMME D'IDENTIFICATION DES GÈNES

L'identification des gènes dans les textes pose de nombreux problèmes, de nature assez diverse et d'importance plus ou moins grande. Par ***identification***, nous entendons à la fois reconnaissance des gènes présents et, pour chaque gène en présence, des occurrences dans le texte où il y est fait référence.

Nous proposons ici un inventaire structuré de ces difficultés. Chaque difficulté est illustrée par des exemples issus de *Medline*. Cette partie a donc pour but de définir précisément en quoi consiste la tache d'identification des gènes dans les textes. Elle constitue un cahier des charges, qui permettra de justifier le schéma du logiciel que nous avons conçu.

Pour chaque difficulté nous donnons un aperçu de la solution que nous proposons, mais les détails techniques de la mise en œuvre seront expliqués dans l'exposé du fonctionnement du logiciel section Chapitre 2I.

A. MÉTHODOLOGIE

Cet inventaire a été rendu possible grâce à une annotation experte des textes. Cette annotation est très précise puisque chaque référence à un gène est identifiée et interprétée comme illustré dans l'exemple ci-dessous.

Exemple 4 Annotation des phrases
Chaque phrase est annotée de façon précise. Ici les termes soulignés *pct*, *hh* et *wg* sont repérés et interprétés comme faisant références aux gènes *patched (pct)*, *hedgehog (hh)* et *wingless (wg)*.

Here we present further evidence that ptc and hh encode components of a signal transduction pathway that regulate the expression of wg transcription following its activation by pair-rule genes.

Ces annotations ont été saisies par un spécialiste du domaine dans une base de données. Un ensemble de 112 résumés, que nous appellerons l'***échantillon A***, a ainsi été complètement annoté. Cela nous a permis de faire un inventaire précis des difficultés rencontrées pour l'identification des gènes dans les textes et de quantifier l'importance de chacune d'elle.

En effet, pour chaque difficulté nous présentons une liste d'exemples qui est exhaustive. Ainsi, la taille du tableau d'exemples indique l'importance de la difficulté.

Comme nous allons le voir, l'identification des gènes dans les textes est un problème complexe. Les règles à appliquer seront donc nombreuses et chaque règle pourra recevoir des exceptions. Voyons maintenant chaque type de difficulté.

B. COMPLEXITÉ DE LA NOMENCLATURE

1. Règles de désignation des gènes pour la drosophile

La mutation d'un gène donné se manifeste généralement par l'apparition d'une caractéristique physique (un phénotype) qui est le plus souvent une anomalie. Souvent le gène prend le nom de ce handicap ou un nom qui l'évoque. Par exemple, le dysfonctionnement du gène *white* se manifeste par une dépigmentation des yeux de l'individu. Les noms peuvent aussi être composés comme par exemple *Suppressor of Hairless* qui est un gène qui inhibe l'expression du gène *Hairless*. Les découvreurs de gènes doivent faire preuve de beaucoup d'imagination pour trouver des noms qui ne sont pas déjà pris et choisissent des noms originaux pour décrire les phénotypes et donc les gènes auxquels ils ont à faire. Voici quelques exemples ci-dessous.

Tableau 5 Exemples de nom de gène
Les biologistes n'utilisent pas de noms de code mais laissent libre cours à leur imagination pour décrire les individus mutants.

Gène (traduction littérale)
Hairless (chauve)
gypsy (bohémien)
hedgehog (hérisson)
gooseberry (groseille à maquerau)

On voit que les gènes peuvent avoir des noms très divers. Il n'y a pas de nom de code avec un format particulier qui permettrait de les reconnaître dans un texte sans avoir à les connaître par avance. Ainsi, on se voit dans l'obligation d'utiliser des lexiques, c'est à dire des listes de termes à rechercher dans les textes.

Voyons maintenant pourquoi l'utilisation d'un lexique non structuré ne convient pas, et pourquoi il est nécessaire d'avoir en sa possession un véritable dictionnaire des gènes.

2. Existence de plusieurs termes pour désigner un seul gène

La tâche d'identification des gènes dans les textes est compliquée par le fait qu'un même gène peut être désigné de plusieurs façons. Il n'est pas rare de voir un auteur utiliser plusieurs noms pour un même gène dans le même résumé ou dans la même phrase. Voici un exemple ci-dessous.

Exemple 5 Plusieurs termes pour désigner un seul gène.
Les mots soulignés, *wg* et *wingless*, désignent le même gène.

The segment polarity gene wingless (wg) is expressed in a complex pattern during embryogenesis suggesting that it plays multiple roles in the development of the embryo.The best characterized of these is its role in cell pattening in each parasegment, a process that requires the activity of other segment polarity genes including patched (ptc) and hedgehog (hh). Here we present further evidence that ptc and hh encode components of a signal transduction pathway that regulate the expression of wg transcription following its activation by pair-rule genes. We also show that most other aspects of wg expression are independent of this regulatory network.

Nous voyons dans cet exemple l'utilisation d'abréviations pour les gènes. Ce sont les ***symboles*** des gènes. Dans une terminologie alternative, ils sont appelés nom abrégé. Le nom standard des gènes est par opposition appelé ***nom complet*** ou parfois nom développé.

L'usage veut que l'on précise la terminologie que l'on emploiera dans le résumé, en écrivant le *nom complet* puis le *symbole* entre parenthèses ; puis que l'on utilise par la suite le *symbole*. C'est ce qui est fait dans l'exemple en page 54.

Malheureusement, s'il est vrai qu'un auteur n'utilise qu'un seul *symbole* et un seul *nom complet*, ce qui ne fait que deux noms en tout, il se peut qu'un autre auteur utilise lui un autre couple *symbole-nom complet*. Nous utilisons les données contenues dans une base de données qui fait référence en la matière pour choisir un *nom complet* et un *symbole* : *Flybase*. Tout autre nom sera considéré comme un ***nom synonyme***.

La présence de plus d'un nom pour un seul gène va nous obliger à utiliser un dictionnaire, c'est à dire un lexique plus structuré qu'une simple liste de termes. Chaque terme qui peut être employé dans un texte pour désigner un gène sera appelé un ***label***. A chaque gène on peut associer un certain nombre de labels. Nous dirons que cela constitue les ***définitions*** du gène.

Nous pouvons nous représenter le dictionnaire comme étant constitué d'entrées qui décrivent chacune un gène. Suivent ensuite les définitions qui donnent chacune un label possible. Chaque définition est d'un type donné : *symbole*, *nom complet* ou *nom synonyme*. Ceci est illustré dans le tableau 6.

Tableau 6 Un gène et ses définitions.
Chaque définition donne un label possible pour désigner le gène. Elle peut être de trois types : symbole, *nom complet* ou *nom synonyme*.

Le gène	Les définitions
wingless (wg)	
	wg, symbole
	wingless, nom complet
	Spd, synonyme
	spade, synonyme
	fg, synonyme
	flag, synonyme
	Sp, synonyme
	Sternopleural, synonyme
	Br, synonyme
	Bristled, synonyme
	int-1, synonyme
	Dint-1, synonyme
	Dm-1, synonyme
	l(2)wg, synonyme

Par construction, les définitions d'un gène utilisent des labels qui sont tous distincts. En revanche, un label peut appartenir à plusieurs définitions. Ce qui veut dire que plusieurs gènes peuvent avoir en commun un même label. Nous dirons dans ce cas que les définitions sont en conflit. Nous reviendrons plus tard sur cette difficulté, mais nous pouvons déjà noter l'utilité de l'utilisation d'une base de données pour représenter de façon efficace la réalité de la terminologie.

Pour situer l'importance relative des différents type de nom, nous donnons dans le tableau suivant les résultats d'une statistique obtenue sur l'échantillon A.

Tableau 7 Importance relative de chaque type de définition
Cette statistique est faite à partir de l'annotation manuelle de l'échantillon A. La fréquence indique le nombre d'occurrence dans les textes de définition d'un type donné.

Type	Fréquence	Proportion
Symbole	539	42%
Nom Complet	375	30%
Synonyme	360	28%
Total	*1274*	*100%*

On constate qu'aucune catégorie n'est négligeable relativement aux deux autres.

Les données utilisées pour construire le dictionnaire ont été extraites d'une base de données très complète pour la nomenclature des gènes de la drosophile : ***Flybase***.

3. Importance de la casse pour désigner un gène

La ***casse***, c'est à dire la caractéristique minuscule ou majuscule d'un caractère, a de l'importance. Le tableau 8 donne des exemples de labels effectivement rencontrés dans les textes pour lesquels il y aurait eu ambiguïté si nous ne disposions pas d'un système de reconnaissance qui fasse la différence entre les majuscules et les minuscules.

Tableau 8 Importance de la casse
Une simple interversion des majuscules en minuscule change tout le sens. Les colonnes Label 1 et Label 2 présentent des termes identiques à la casse près, or les gènes désignés sont différents.

Label 1	Gène 1	Label 2	Gène 2
ac	achaete (ac)	Ac	lethal (2) 37Ac (l(2)37Ac)
asx	ascutex (asx)	Asx	Additional sex combs (Asx)
cry	Suppressor of Stellate (Su(Ste))	Cry	Crystallin (Cry)
delta	deltaTrypsin (deltaTry)	Delta	Delta (Dl)
dl	dorsal (dl)	Dl	Delta (Dl)
dl	duplicated legs (dpl) [25]	Dl	Delta (Dl)
hb	hunchback (hb)	HB	HB element (HB)
lip	canoe (cno)	Lip	Lighten up (Lip)
pcl	pepsinogen-like (pcl)	Pcl	Polycomblike (Pcl)
psc	pseudoscute (psc)	Psc	Posterior sex combs (Psc)
Rac	(Rac1)	RAC	(Akt1)
rho	rhomboid (rho)	Rho	Rho1
rho	rhomboid (rho)	Rho	Larval cuticle protein 10 (Lcp10) [26]
ste	Stellate (Ste)	Ste	Suppressor of Stellate (Su(Ste))

Notons que ce tableau, s'il était exhaustif, compterait plus de mille lignes. Nous voyons donc qu'il est nécessaire de disposer d'un système qui sache indexer les termes en prenant en compte la casse. Ce n'est généralement pas le cas des systèmes de gestion électronique de documents. Cela nous a aussi posé des problèmes pour la mise en œuvre de notre base de données sous ***Access***, car ce système de gestion de base de données considère comme égaux des enregistrements (des données) qui ne diffèrent que par la casse.

[25] Les gènes *duplicated legs (dpl)* et *dorsal (dl)* admettent tous deux le label *dl* dans leurs définitions. Cela constitue une contradiction. Nous discuterons de ce type de difficulté dans la section Partie 2Chapitre 1I.E.1

[26] Les gènes *Larval cuticle protein 10 (Lcp10)* et *Rho1* admettent tous deux le label *Rho* dans leurs définitions. Cela constitue une contradiction. Nous discuterons de ce type de difficulté dans la section Partie 2Chapitre 1I.E.1

4. Complexité introduite par la formation de mots composés

Les labels sont utilisés pour former des mots composés. Le tableau 9 donne les exemples que nous avons trouvés lors de l'annotation de l'échantillon A.

Tableau 9 Expressions spécifiques
Des labels entrent dans la composition de mots composés. Ces exemples sont tirés de l'annotation l'*échantillon A*.

Label	Contexte
wingless	- cells
Dfd	- dependent
wingless	- embryos
cry	- males
esc	- mothers
Dorsal	-binding sites
dl	-binding sites
Kr	-binding sites
sna	-binding sites
dl	-binding sites
cyclin A	-cdc2 kinase
H1	-containing 30 nm fibre
actin	-crosslinking protein
PKA	-deficient oocytes
E2F	-dependent transcription
Wingless	-expressing cells
wingless	-expressing cells
copia	-induced allele
cry	-induced meiotic drive
Abdominal-B	-like HOM proteins
dl	-mediated repression
hairy	-mediated repression
en	-mediated repression
hairy	-related bHLH proteins
hairy	-related bHLH proteins encoded by
SD	-specific 4-kb transcript
gurken	-torpedo signaling process

Il est important de repérer ces labels dans les textes malgré l'existence des tirets.

Les premières lignes du tableau font apparaître des espaces après le tiret. Il s'agit d'un formatage particulier propre à *Medline*. Il a probablement été fait pour permettre l'indexation des termes qui suivent le tiret. Quoi qu'il en soit, ce formatage n'est pas systématique comme on peut le constater dans les exemples qui sont donnés.

5. Complexité introduite par l'inclusion des termes les uns dans les autres

a. Inclusion à l'intérieur du dictionnaire des gènes

Une autre difficulté réside dans le fait que les labels s'emboîtent les uns dans les autres à la manière des poupées russes : on utilise le nom d'un gène pour construite le nom d'un autre gène. *Hairless* est un gène, *Supressor of Hairless* un autre gène, le second ayant la particularité d'inhiber l'expression du premier. Nous dirons dans ce cas que le label *Hairless* est ***inclus*** dans le label *Supressor of Hairless*.

Voici un exemple de phrase où cette inclusion pose un problème d'interprétation.

Exemple 6 Inclusion des labels
La première occurrence de *Hairless* (soulignée) ne doit pas être interprétée comme le gène du même nom car elle est incluse dans le label *Suppressor of Hairless* qui désigne un autre gène. En revanche la seconde occurrence de *Hairless* fait bien référence au gène *Hairless (H).*

These results, along with the intermediate SOP phenotype observed in Suppressor of Hairless; Hairless double mutant imaginal discs, suggest that the two genes act antagonistically to commit imaginal disc cells stably to alternative fates.

Le danger consiste à repérer à tort le label inclus sans repérer le label qui l'inclut, où ce qui est plus grave, repérer les deux labels à la fois. La solution passe nécessairement par la tenue d'une table d'inclusion qui consigne chacune d'elles de façon à pouvoir y faire référence lors du repérage des labels dans les textes. Le tableau 10 en donne un extrait.

Tableau 10 Table d'inclusion des labels
Les labels sont inclus les uns dans les autres. Le Label 1 est inclue dans le Label 2, mais ils désignent des gènes différents.

Label 1	Label 2	Gène 1	Gène 2
fu	Su(fu)	fused (fu)	Suppressor of fused (Su(fu))
H	Su(H)	Hairless (H)	Suppressor of Hairless (Su(H))
Hairless	Suppressor of Hairless	Hairless (H)	Suppressor of Hairless (Su(H))
knirps	knirps-related	knirps (kni)	knirps-like (knrl)
Pc	E(Pc)	Polycomb (Pc)	Enhancer of Polycomb (E(Pc))
P-element	P-element somatic inhibitor	P element (P-element)	P-element somatic inhibitor (Psi)
pn	K-pn	prune (pn)	abnormal wing discs (awd)
scute	lethal of scute	scute (sc)	lethal of scute (l(1)sc)
Scute	Lethal of Scute	scute (sc)	lethal of scute (l(1)sc)
scute	lethal-of- scute	scute (sc)	lethal of scute (l(1)sc)
sev	E(sev)3A	sevenless (sev)	Heat shock protein 83 (Hsp83)
sev	E(sev)3B	sevenless (sev)	(Cdc37)
sevenless	bride of sevenless	sevenless (sev)	bride of sevenless (boss)
stoned	Suppressor of stoned	stoned B (stnB)	Suppressor of stoned (Su(stn))
stoned	Suppressor of stoned	stoned A (stnA)	Suppressor of stoned (Su(stn))
tolloid	tolloid-related-1	tolloid (tld)	tolkin (tok)
tra	tra-2	transformer (tra)	transformer 2 (tra2)
transformer	transformer 2	transformer (tra)	transformer 2 (tra2)
Trithorax	Trithorax-like	torso (tor)	Trithorax-like (Trl)
white	Zeste-white 3	white (w)	shaggy (sgg)
z	E(z)	zeste (z)	Enhancer of zeste (E(z))
z	E(z)1	zeste (z)	Enhancer of zeste (E(z))
z	Su(z)2	zeste (z)	Suppressor of zeste 2 (Su(z)2)
z	Su(z)2(1)	zeste (z)	Suppressor of zeste 2 (Su(z)2)
zeste	Enhancer of zeste	zeste (z)	Enhancer of zeste (E(z))

Notons qu'il se peut que l'inclusion se fasse pour des labels utilisés pour définir le même gène. Dans l'extrait que nous avons présenté, les inclusions sont relatives à des gènes distincts. Notons que le tableau, s'il était exhaustif, compterait un peu plus de 4000 lignes.

b. Inclusion des labels dans des termes de biologie

Le même phénomène d'inclusion a lieu aussi entre les labels et des expressions de biologie qui ne font pas référence à des gènes. Voici un exemple en page 59.

Exemple 7 Inclusion des labels dans des termes de biologie
Le terme souligné *Pc* ne fait pas référence au gène *Polycomb (Pc)* mais fait partie de l'expression *Pc group* qui désigne un complexe de protéine.

We have examined the pattern of expression of the Drosophila segment polarity gene, engrailed (en), in embryos mutant for several different members of the Pc group.

L'inclusion peut avoir lieu avec des termes qui font référence à des *complexes de gènes* ou des *complexes de protéines*. Le tableau 11 fournit des exemples constitués à partir de complexes que nous avons trouvés dans les textes.

Tableau 11 Confusion avec des complexes de gènes ou de protéine
Le nom des complexes de gènes ou de protéine peuvent être formés à partir de nom de gène. Ici les labels de la première colonne sont inclus dans les labels de la seconde colonne. Il est nécessaire de prendre en compte certains complexes pour reconnaître correctement les gènes.

Label du gène	Label du complexe	Gène	Complexe
achaete	achaete-scute complex	achaete (ac)	Achaete-scute Complex (ASC)
Antennapedia	Antennapedia complex	Antennapedia (Antp)	Antennapedia complex (ANT-C)
Enhancer of split	Enhancer of split complex	Enhancer of split (E(spl))	Enhancer of split complex
Pc	Polycomb (Pc) group	Polycomb (Pc)	Polycomb group (Pc-G)
Pc	Pc-G	Polycomb (Pc)	Polycomb group (Pc-G)
Pc	Pc group	Polycomb (Pc)	Polycomb group (Pc-G)
Polycomb	Polycomb (Pc) group	Polycomb (Pc)	Polycomb group (Pc-G)
Polycomb	Polycomb group	Polycomb (Pc)	Polycomb group (Pc-G)
Shaker	Shaker complex	Shaker (Sh)	Shaker complex (ShC)
scute	Achaete-scute Complex	scute (sc)	Achaete-scute Complex (ASC)
scute	achaete-scute	scute (sc)	Achaete-scute Complex (ASC)

Il se peut aussi que l'inclusion ait lieu avec des termes anatomiques ou plus généralement avec des termes appartenant à l'univers de la biologie. Le tableau 12 fournit des exemples que nous avons trouvés en annotant les textes.

Tableau 12 Confusion avec des termes de génétique ou d'anatomie
Les labels listés sont inclus dans des termes de génétique, d'anatomie ou autre.

Label inclus	Expression prêtant à confusion
AR	adaptive response (AR)
cell	cell-cell interaction
arm	chromosome arm, C-terminal arm, N-terminal arm
disrupted	disrupted polarity
dorsal	dorsal side, dorsal vessel, dorsal cell, dorsal closure, dorsal half, dorsal ectoderm, dorsal follicle, dorsal fate, dorsal pattern, dorsal-ventral, dorsal epidermis, dorsal midline
G	G phase
disc	entire disc, imaginal disc, wing disc
mis	mis expression
furrow	morphogenetic furrow, ventral furrow
P element	P element transformation, P element mediated transformation
pupal	pupal stage
ring	ring canal
patch	patch of
ref	see ref
slight	slight effect
side	ventral side

Dans tous les cas, nous voyons qu'il y a confusion possible de certains labels avec des expressions qui ne désignent pas des gènes. Repérer l'inclusion permet de lever l'ambiguïté qui existe au départ. Il est nécessaire de disposer d'un lexique complémentaire pour reconnaître correctement les gènes.

Lors de l'annotation, nous avons repéré ces termes de biologie qui incluent des labels en leur sein et nous avons créé de nouvelles entrées dans le dictionnaire pour les prendre en compte lors de l'identification automatique des gènes dans les textes. Cependant, comme ce ne sont pas à proprement parler des gènes, nous avons créé de nouvelles rubriques. Ces rubriques sont intitulées ***complexe de gènes***, ***complexe de protéines*** et ***terme spécifique***. Cette dernière rubrique correspond aux exemples du tableau 12.

6. Complexité introduite par l'existence des allèles

Les allèles d'un gène sont les différents états que peuvent prendre un même gène chez un individu. Ce sont des variantes sur un même gène. Le dictionnaire que nous avons extrait de *Flybase* ne comportait pas initialement d'allèle. Cependant lors de l'annotation des textes nous avons rencontré des références à des allèles. Le tableau 13 en fait la liste.

Tableau 13 Les allèles
Les gènes admettent des allèles. La colonne de droite donne le gène de référence pour l'allèle de la colonne de gauche. Ces exemples sont issus de l'annotation des textes.

Allèle	Gène
AntpNs	Antennapedia (Antp)
enhancer of rudimentaryp1 (e(r)p1)	enhancer of rudimentary (e(r))
E(z)1	Enhancer of zeste (E(z))
Psc1	Posterior sex combs (Psc)
Sce1	Sex combs extra (Sce)
ScmD1	Sex combs on midleg (Scm)
Su(z)2(1)	Suppressor of zeste 2 (Su(z)2)
tolloid-related-1	tolkin (tok)
white-apricot (wa)	white (w)
white-blood (wbl)	white (w)
wnt-1	wingless (wg)

Pour parler de la relation qu'entretient un gène avec ses allèles, nous définissons la notion de ***gène de référence***. Le gène de référence d'un allèle est le gène associé à cet allèle. Le gène de référence d'un gène n'est autre que lui-même.

Il est important de savoir reconnaître les allèles pour deux raisons.

D'une part, les noms d'allèles sont souvent composés à partir de noms de gène. Les labels d'allèles participent donc au problème qui a été expliqué dans la section Partie 2Chapitre 1I.B.5.

D'autre part, quand un auteur décrit une interaction en citant un ou plusieurs allèles, il faut comprendre cette interaction comme ayant lieu entre le ou les gènes associés. Il est donc nécessaire de reconnaître les allèles dans les textes et de faire le lien entre l'allèle et le gène auquel il est associé. Nous avons donc introduit une rubrique ***allèle*** dans notre dictionnaire et nous avons complété notre dictionnaire avec les allèles rencontré dans les textes. Un lien entre l'allèle et son gène de référence a aussi été établi.

C. AMBIGUÏTÉ DES LABELS

Nous arrivons à la difficulté sans doute la plus importante. Les labels sont parfois ***ambigus***, c'est à dire qu'ils peuvent faire référence à tout autre chose que des gènes. Nous avons distingué deux catégories de label ambigu selon la gravité de la situation.

1. Les labels qui sont des *mots vides*

Cette première catégorie de label ambigu correspond à des mots extrêmement courants de l'anglais comme *if* ou *for* qui malheureusement désignent des gènes. Le tableau 14 donne la liste de ces termes.

Tableau 14 *Labels* et *mots vides*
Les labels présentés prêtent à confusion avec des *mots vides*.

Label	Gène
an	ancon (an)
as	ascute (as)
at	arctus oculus (at)
be	tumor(3)be (tu(3)be)
by	blistery (by)
can	cannonball (can)
did	diminished discs (did)
do	pale ocelli (po)
for	foraging (for)
her	hermaphrodite (her)
how	held out wings (how)
if	inflated (if)
in	inturned (in)
me	focal melanosis (me)
none[27]	glass (gl)
not	non-stop (not)
or	orange (or)
per	period (per)
she	sherry (she)
so	sine oculis (so)
up	upheld (up)
us	undersized (us)
we	wee (we)
who	held out wings (how)[28]
with	with trident (with)

Ces mots sont appelés ***mots vides*** (*stop word* en anglais) en recherche documentaire. Cette appellation provient du fait que ces mots à eux seul ne renferment pas de sens. C'est à dire que leur présence ou absence dans un texte donné ne permet pas de savoir quoi que ce soit sur ce texte quant à son sens. Ils ne sont donc jamais utilisés dans les index. Nous avons employé une liste de *mots vides* établie pour la mise au point d'un système d'indexation de texte en anglais. Nous avons trouvé parmi les labels de notre dictionnaire un certain nombre de termes qui appartiennent à cette liste. Nous voyons que la liste des membres de cette première catégorie de labels ambigus a été établie avant toute expérience ; ce qui ne sera pas le cas de la deuxième liste qui elle sera établie à la lecture des textes, au fur et à mesure de la rencontre avec des labels ambigus. Certains termes ont pu quand même changer de catégorie, quand nous nous sommes aperçus qu'ils n'étaient pas toujours aussi largement répandus dans les textes.

Les occurrences de ces termes sont trop nombreuses pour que nous puissions les indexer systématiquement. Cela aboutirait à une surcharge de la base de données. Il est d'ailleurs d'usage de ne pas les inclure dans les index en partie pour cette raison.

Néanmoins nous verrons que le contexte permet dans certain cas de les prendre en compte lors de l'identification des gènes dans les textes. Retenons simplement pour l'instant que la présence à elle seule d'un de ces labels ne peut être interprété comme une référence à un gène.

---

[27] l'appellation *none* provient de l'abréviation du synonyme *no-ocelli--narrow-eyes*

[28] Le gène *held out wings (how)* a bien comme définition synonyme le label *who*. Le gène compte en effet *wings held out* dans ces définitions, ce qui explique la présence de *who*.

Les labels que nous avons présentés dans le tableau 14 se confondent exactement avec des *mots vides*. D'autres labels ne s'en distinguent que par la casse. Ils sont présentés dans le tableau 15.

Tableau 15 *Mots vides* et différence de casse
Les labels présentés prêtent à confusion avec des *mots vides*, mais ils s'en distinguent par la casse.

Label	Gène	Remarque
And	Androcam (And)	
At	Attenuated (At)	
Be	lethal (2) 37Be (l(2)37Be)	
Can	Calcineurin B (CanB)	Can est un label commun à deux gènes
Can	Calcineurin A1 (CanA1)	idem
Co	Notch (N)	Co provient du synonyme Confluens
Had	beta Hydroxy acid dehydrogenase (Had)	
Is	Isis (Is)	
Low	Lightener of white (Low)	
Me	Moire (Me)	
Off	Off	
On	Open (On)	
Re	Re	
To	Superoxide dismutase (Sod)	To provient du synonyme Tetrazolium oxidase
Ve	veinlet (ve)	
We	Washed eye (We)	

Ces labels sont recherchés dans les textes car le système prend en compte la différence de casse. Cependant quand le mot en question se trouve en première position dans la phrase, alors il y a de fortes chances pour qu'il s'agisse en fait du mot vide correspondant. Dans ce cas la reconnaissance ne se fait pas.

2. Les labels qui prêtent à confusion avec des termes d'anglais assez courants

Cette deuxième catégorie de labels ambigus comporte des termes dont l'ambiguïté est moins sévère. Ce sont des termes de la langue anglaise, mais ce ne sont pas des mots-outils, des mots à tout faire comme pour la première catégorie.

a. Les labels fortement ambigus

Certains labels sont, dans le contexte des textes que nous étudions, fortement ambigus. Le tableau 33[29] en fait l'inventaire.

Ces termes sont trop ambigus pour que l'on puisse avoir totalement confiance quand on les rencontre dans les textes. Nous verrons à la section F qu'une utilisation du contexte permet de résoudre le problème.

b. Les labels qui dans le contexte de la génétique sont moins ambigus qu'ils ne semblent

Certains termes, bien que faisant partie du dictionnaire, ne sont pas aussi ambigus qu'il y paraît, et pourront être utilisés pour identifier les gènes dans les textes. C'est le cas par exemple de *hedgeog*, qui bien qu'ayant un autre sens que celui d'un gène (hérisson), devra être

[29] Les tableaux longs sont placés en fin de partie de niveau deux (numérotées A, B, C, etc.)

interprété comme une référence au gène *hedgeog (hh)* car il est peu probable que l'on parle de hérisson dans un texte de génétique de la drosophile. Le tableau 16 fait l'inventaire de ces termes.

Tableau 16 Labels peu ambigus
Les labels présentés sont a priori ambigus, mais pas dans le contexte de la génétique.

Label	Gène	Remarques
cap	Calphotin (Cpn)	Signifie chapeau, terme rare
cap	capon (cap)	idem
cap	Chromosome-associated protein (Cap)	idem
Deformed	Deformed (Dfd)	ne se confond pas avec deformed
giant	giant (gt)	signifie géant, terme rare
HAD	beta Hydroxy acid dehydrogenase (Had)	ne se confond pas avec had
hedgehog	hedgehog (hh)	Signifie hérisson, terme rare
ltd	lightoid (ltd)	ltd est l'abréviation de limited, terme rare
ME	Malic enzyme (Men)	ne se confond pas avec me
mr	morula (mr)	ne se confond pas avec Mr
rough	rough (ro)	signifie rugueux, terme rare
stranded	stranded (sand)	signifie échouer, terme rare
suffix	suffix element (suffix)	
thick	thick (tk)	signifie gros, terme rare
thin	thin (tn)	signifie mince, terme rare
tube	tube (tub)	
weak	weak (wk)	signifie faible, terme rare

Nous constatons que, dans certains cas, c'est la différence de casse entre le label et le terme d'anglais qui permet de lever l'ambiguïté.

c. Les labels faiblement ambigus.

Pour d'autres termes, nous n'avons pas beaucoup d'exemples d'occurrence dans les textes. Ainsi, il s'agit de termes qui sont rares à la fois dans leur acception de label et dans une autre acception. Cependant, ils paraissent assez peu ambigus et ils seront donc utilisés pour l'identification des gènes dans les textes. Le tableau 34 fait l'inventaire de ces termes.

d. Les labels ambigus mais très importants

Les gènes *dorsal (dl)* et *armadillo (arm)* sont très importants en génétique de la drosophile et sont souvent cités par les auteurs. Ne pas les reconnaître serait donc grave à la fois du point de vue de la biologie et du point de vue des performances attendues du système. Une stratégie de désambiguïsation a donc été mise en œuvre. Elle est basée sur la présence d'un contexte qui dans le cas où il ne s'agirait pas d'un gène, va donner des indices qui permettent de lever l'ambiguïté. Ces termes de désambiguïsation, déjà donnés dans le tableau 12, sont listés à nouveau dans le tableau 17.

Tableau 17 Labels désambiguïsés
Les termes *dorsal* et *arm* sont pris en compte grâce à une technique de désambiguïsation. Le label est interprété comme un gène sauf si c'est le terme de désambiguïsation qui est reconnu.

Label	Terme de désambiguïsation
arm	chromosome arm
arm	N-terminal arm
arm	C-terminal arm
dorsal	dorsal side
dorsal	dorsal vessel
dorsal	dorsal-specific
dorsal	dorsal cell
dorsal	dorsal closure
dorsal	dorsal half
dorsal	dorsal ectoderm
dorsal	dorsal follicle
dorsal	dorsal fate
dorsal	dorsal cells
dorsal	dorsal or ventral
dorsal	dorsal pattern
dorsal	dorsal-ventral
dorsal	dorsal epidermis
dorsal	dorsal midline

3. Les labels qui prêtent à confusion avec des gènes de mammifères.

Dans les textes que nous analysons, il est parfois question de gènes d'autres espèces biologiques comme dans l'exemple ci-dessous.

Exemple 8 Confusion possible avec des gènes de mammifères
Des gènes de mammifères ont les noms qui se confondent avec ceux de la drosophile. Ici les occurrences soulignées de *E2F* font référence à un gène humain, alors que *E2F* est un synonyme de *E2F transcription factor (E2f)* d'après *Flybase.*

The temporal activation of E2F transcriptional activity appears to be an important component of the mechanisms that prepare mammalian cells for DNA replication. Regulation of E2F activity appears to be a highly complex process, and the dissection of the E2F pathway will be greatly facilitated by the ability to use genetic approaches. We report the isolation of two Drosophila genes that can stimulate E2F-dependent transcription in Drosophila cells. One of these genes, dE2F, contains three domains that are highly conserved in the human homologs E2F-1, E2F-2, and E2F-3. Interestingly, one of these domains is highly homologous to the retinoblastoma protein (RB)-binding sequences of human E2F genes. The other gene, dDP, is closely related to the human DP-1 and DP-2 genes. We demonstrate that dDP and dE2F interact and cooperate to give sequence- specific DNA binding and optimal trans-activation. These features suggest that endogenous Drosophila E2F, like human E2F, may be composed of heterodimers and may be regulated by RB-like proteins. The isolation of these genes will provide important reagents for the genetic analysis of the E2F pathway.

Les auteurs font référence à des gènes d'autres espèces pour donner des informations sur l'homologie ou, d'une façon plus générale, sur les ressemblances dans les propriétés ou fonctions des gènes en question avec des gènes de la drosophile.

Le problème se rencontre 13 fois dans l'échantillon A. Le tableau 18 fournit les phrases concernées.

Tableau 18 Occurrence de gène de mammifère
La colonne de droite donne le label qui prête à confusion. Une référence à un gène de mammifère ne doit pas être interprétée comme une référence à un gène de drosophile.

Phrase	Label
We show here that btd is expressed in a stripe covering the head analgen of the segments affected in btd lack-of-function mutants and that btd encodes a zinc-finger-type transcription factor with sequence and functional similarity to the prototype mammalian transcription factor Sp1.	Sp1
When expressed in the spatial pattern of btd, a transgene providing Sp1 activity can support development of the mandibular segment in the head of btd mutant embryos.	Sp1
This Musca protein, designated Musca PRI, changes its pI upon illumination in vivo.	PRI et PRIs
Rabbit antibodies raised against Musca PRI, against bovine arrestin, and against a synthetic peptide based on the Drosophila PRI sequence stained the Drosophila and Musca PRIs specifically on 1 and 2-dimensional Western immunoblots.	PRI et PRIs
Both Drosophila and Musca PRIs incorporated 32P-radioactivity from gamma-32P-ATP in cell-free homogenates of retinas.	PRI et PRIs
Partial peptide digestions of Drosophila and Musca PRIs revealed similarity between these proteins.	PRI et PRIs
Mutations in the Drosophila gene extradenticle (exd), a homologue of the human proto-oncogene pbx1, cause homeotic transformations by altering the morphological consequences of homeotic selector gene activity. exd has been proposed to act by contributing to the specificity of selector homeodomain proteins for their downstream targets.	pbx1
The Drosophila protein Dorsal (which, like the human protein NF-kappa B3, is a member of the Rel family of transcriptional activators) activates the twist gene and represses the zen gene in the ventral region of early embryos.	Rel

Nous constatons que dans la plupart des cas, il suffit de compléter le dictionnaire en créant une rubrique pour les gènes de mammifères. Nous n'avons pas complété cette rubrique à l'aide de bases de donnée de génétique de la souris, de l'homme ou des mammifères. Nous avons simplement, et à titre expérimental, complété cette rubrique avec les exemples que nous avons trouvés au cours de l'annotation. Le tableau 19 liste les données ainsi introduites.

Tableau 19 Gène de mammifère : extrait du dictionnaire
Des gènes de mammifères qui ont été rajoutés au dictionnaire des gènes. Seul les cas effectivement trouvés dans les textes sont concernés. Ces informations ont été introduites par l'annotateur à titre expérimental, pour prouver la capacité du système à correctement distinguer entre gène de drosophile et gène d'autres espèces.

Gène	Label
E2F-1	E2F-1
E2F-1	human E2F
E2F-2	E2F-2
E2F-2	human E2F
E2F-3	E2F-3
E2F-3	human E2F
human proto-oncogene pbx1	human proto-oncogene pbx1
mammalian transcription factor Sp1	mammalian transcription factor Sp1
mammalian transcription factor Sp1	Sp1
Musca PRI	Musca PRI
Musca PRI	Musca PRIs

D. IMPRÉCISION DANS LA TERMINOLOGIE

1. Les termes qui ne décrivent pas un gène précis mais qui peuvent désigner plusieurs gènes

Dans le dictionnaire, certains labels participent à plusieurs définitions. Nous dirons alors que le label est ***imprécis***. Dans l'exemple ci-dessous, le label *hsp70* est reconnu car il fait bien partie du dictionnaire, mais le dictionnaire fournit non pas un, mais quatre candidats possibles pour ce label.

Exemple 9 Imprécision dans la terminologie
Le texte peut ne pas préciser exactement de quel gène il s'agit. Ici l'auteur en notant *hsp70* (souligné) ne précise pas s'il s'agit de *Heat-shock-protein 70Aa, 70Ab, 70Ba, 70Bb* ou *70Bc*.

Immunopurified TFIID produces a large DNase I footprint over the hsp70, hsp26, and histone H3 promoters of Drosophila.

L'auteur n'est pas assez précis par rapport au dictionnaire que nous avons. Ce phénomène est d'autant plus préoccupant qu'une interaction est décrite, mais l'auteur ne dit pas exactement avec quel gène. L'information qu'il fournit est vraiment relative à ce que nous appellerons une ***collection*** de gènes et pas à un gène particulier. Ce phénomène est assez courant. Lors de l'annotation experte des textes, nous avons répertorié une série de labels qui présentent cette caractéristique. Ils sont présentés dans le tableau 35.

Pour prendre en compte cette imprécision et annoter les textes malgré tout, nous avons créé de nouvelles entrées dans le dictionnaire. Ainsi, par exemple, nous avons créé un nouvel objet que nous avons nommé *hsp70* et qui admet comme label *hsp70*. Parallèlement, la phrase de l'exemple ci-dessus sera annotée en signalant que l'occurrence de *hsp70* doit être comprise comme une référence à l'objet nouvellement créé dans le dictionnaire. Ce dispositif nous permet d'annoter le plus fidèlement possible les phrases. L'auteur ne fait pas référence à un des éléments de la collection, ni même à chacun des éléments de la collection. Il fait référence à la *collection*, qui n'est ni réductible à un élément particulier, ni à son ensemble. Cependant des liens ont été créés entre les entités nouvellement créés (de type *collection*) et les membres de la *collection* (de type gène).

Chaque élément créé (de type collection) est classé dans l'une des trois catégories : *famille de gènes*, *famille de protéines* ou *complexe de gènes*.

2. Les variations orthographiques

Certaine orthographies sont absentes du dictionnaire fournis par *Flybase*.

a. Inventaire des orthographies absentes de Flybase

Au cours de l'annotation des textes nous avons relevé toutes les variations orthographiques non-répertoriées dans *Flybase*. Le tableau 36 en dresse l'inventaire.

Dans la plus part des cas, il existe dans *Flybase* une définition approchante. Nous dirons que la nouvelle définition est une ***variante*** de l'ancienne définition et que les labels sont liés par une ***relation de transformation***. Certaines transformations sont automatisable. Nous dirons que les variantes correspondantes sont ***prévues***.

b. Les variantes prévues

Il existe quatre types de relation de transformation qui sont prise en charge par le système.

Le type le plus important de relation de transformation est le type *première lettre en majuscule*. Les définitions qui sont concernés sont listées dans le tableau 20.

Tableau 20 Transformation de type première lettre en majuscule
Le label 1 était présent dans les définitions *Flybase* du gène, mais pas le label 2. Le label 2 a été utilisé au moins une fois pour désigner le gène.

Label 1	Transformation	Label 2	Gène
achaete	1ière lettre en majuscule	Achaete	achaete (ac)
armadillo	1ière lettre en majuscule	Armadillo	armadillo (arm)
bicoid	1ière lettre en majuscule	Bicoid	bicoid (bcd)
cactus	1ière lettre en majuscule	Cactus	cactus (cact)
daughterless	1ière lettre en majuscule	Daughterless	daughterless (da)
dorsal	1ière lettre en majuscule	Dorsal	dorsal (dl)
hb	1ière lettre en majuscule	Hb	hunchback (hb)
hunchback	1ière lettre en majuscule	Hunchback	hunchback (hb)
nos	1ière lettre en majuscule	Nos	nanos (nos)
pelle	1ière lettre en majuscule	Pelle	pelle (pll)
runt	1ière lettre en majuscule	Runt	runt (run)
scute	1ière lettre en majuscule	Scute	scute (sc)
sevenless	1ière lettre en majuscule	Sevenless	sevenless (sev)
sry delta	1ière lettre en majuscule	Sry delta	Serendipity delta (Sry-delta)
torso	1ière lettre en majuscule	Torso	torso (tor)
trithorax	1ière lettre en majuscule	Trithorax	trithorax (trx)
tube	1ière lettre en majuscule	Tube	tube (tub)
wingless	1ière lettre en majuscule	Wingless	wingless (wg)

Vient ensuite le cas de relation de transformation de type *tout en majuscule*. Le tableau 21 donne les définitions concernées effectivement reconnues par l'annotateur dans l'échantillon A.

Tableau 21 Transformation de type tout en majuscule
Le label 1 était présent dans les définitions *Flybase* du gène, mais pas le label 2. Le label 2 a été interprété au moins une fois comme une manifestation du gène par l'annotateur.

Label 1	Transformation	Label 2	Gène
Antp	Tout en majuscules	ANTP	Antennapedia (Antp)
Psi	Tout en majuscules	PSI	P-element somatic inhibitor (Psi)
tra	Tout en majuscules	TRA	transformer (tra)
UBx	Tout en majuscules	UBX	Ultrabithorax (Ubx)
Ubx	Tout en majuscules	UBX	Ultrabithorax (Ubx)
antp	Tout en majuscules	ANTP	Antennapedia (Antp)
dpp	Tout en majuscules	DPP	decapentaplegic (dpp)
scw	Tout en majuscules	SCW	screw (scw)
tolloid	Tout en majuscules	TOLLOID	tolloid (tld)
ubx	Tout en majuscules	UBX	Ultrabithorax (Ubx)

Notons que le label transformé correspond souvent à la protéine synthétisée par le gène. C'est la raison pour laquelle nous avons choisi de définir par défaut le type de ces définitions à ***protéine***.

Vient ensuite le cas de relation de transformation de type *espace transformé en tiret*. Le tableau 22 donne les définitions concernées effectivement reconnues par l'annotateur dans l'échantillon A.

Tableau 22 Transformation de type espace transformé en tiret
Le label 1 était présent dans les définitions *Flybase* du gène, mais pas le label 2. Le label 2 a été interprété au moins une fois comme une manifestation du gène par l'annotateur.

Label 1	transformation	Label 2	Gène
Abdominal B	espace -> tiret	Abdominal-B	Abdominal B (Abd-B)
Bicaudal D	espace -> tiret	Bicaudal-D	Bicaudal D (BicD)
Sex lethal	espace -> tiret	Sex-lethal	Sex lethal (Sxl)
abdominal A	espace -> tiret	abdominal-A	abdominal A (abd-A)
even skipped	espace -> tiret	even-skipped	even skipped (eve)
gooseberry distal	espace -> tiret	gooseberry-distal	gooseberry distal (gsb-d)
gooseberry proximal	espace -> tiret	gooseberry-proximal	gooseberry proximal (gsb-p)

Vient enfin le cas de relation de transformation de type *tout en minuscule* ou *tiret transformé en espace*. Le tableau 23 donne les définitions concernées effectivement reconnues par l'annotateur dans l'échantillon A.

Tableau 23 Transformation de type tout en minuscule ou tiret transformé en espace
Le label 1 était présent dans les définitions *Flybase* du gène, mais pas le label 2. Le label 2 a été utilisé au moins une fois dans les textes annotés pour désigner le gène.

Label 1	transformation	Label 2	Gène
troponin-I	tiret -> espace	troponin I	wings up A (wupA)
Sry-delta	tiret -> espace	Sry delta	Serendipity delta (Sry-delta)
Phosrestin-II	tiret -> espace	Phosrestin II	Arrestin A (Arr1)
Serendipity delta	Tout en minusc.	serendipity delta	Serendipity delta (Sry-delta)
Adducin-like	Tout en minusc.	adducin-like	hu li tai shao (hts)

Bien sûr, en générant automatiquement de nouvelles définitions, on peut créer de nouvelles difficultés. En effet le label transformé peut être ambigu, c'est à dire qu'il se confond avec un terme souvent présent dans les textes mais qui n'a rien à voir avec un quelconque gène. Nous verrons à la section Partie 2Chapitre 1I.F comment cette difficulté peut être résolue par l'utilisation du contexte.

Par ailleurs, il est important de noter que les exemples donnés dans les sections précédentes et en particulier dans la section 5 qui traite de l'ambiguïté des labels, n'ont rien à voir avec ces nouvelles définitions. Autrement dit, les problèmes que nous avons illustrés précédemment n'ont pas été générés par l'introduction automatique de variantes sur les définitions. En effet, nous avons pris soin, dans ces exemples, de n'utiliser que des définitions, soit directement issues de *Flybase*, soit introduites manuellement par l'annotateur.

Pour évaluer l'importance relative de chaque type de relation de transformation nous proposons le tableau 24.

Tableau 24 Importance relative de chaque type de transformation
La colonne *Effectif* donne le nombre de reconnaissance faite par l'annotateur dans l'échantillon A.

Transformation	Effectif
1ière lettre en majuscule	65
Tout en majuscules	23
espace -> tiret	21
tiret -> espace	5
Tout en minuscules	4

c. Les variantes imprévues

Voyons le cas des définitions variantes qui ne sont pas actuellement anticipées par le système que nous proposons. Le tableau 25 répertorie celles que nous avons trouvées.

Tableau 25 Variantes imprévues
Chaque ligne du tableau donne une définition rencontrée au moins une fois dans l'échantillon A. Elles sont toutes absentes du dictionnaire issu de *Flybase*. Ces variantes ne sont pas actuellement prévues par le système. Elles ont été introduites manuellement par l'annotateur.

Label	Gène
abdominal- A	abdominal A (abd-A)
Absent, small or homeotic discs1	absent, small, or homeotic discs 1 (ash1)
AceIJ40	Acetylcholine esterase (Ace)
acetylcholinesterase	Acetylcholine esterase (Ace)
Acetylcholinesterases	Acetylcholine esterase (Ace)
alpha- spectrin	alpha Spectrin (alpha-Spec)
cyclin E	Cyclin E (CycE)
D- Mek	Downstream of raf1 (Dsor1)
D-mekts	Downstream of raf1 (Dsor1)
dorsal switch protein	Dorsal switch protein 1 (Dsp1)
double sex	doublesex (dsx)
EGF-Receptor	EGF receptor (Egfr)
Extra sex combs	extra sexcombs (esc)
Extramacrochaete	extra macrochaetae (emc)
extra-macrochaete	extra macrochaetae (emc)
histone H1	Histone H1 (His1)
histone H3	Histone H3 (His3)
Hsp90	Heat shock protein 83 (Hsp83)
I elements	I element (I-element)
Lethal of Scute	lethal of scute (l(1)sc)
lethal-of- scute	lethal of scute (l(1)sc)
phosrestins I	Arrestin B (Arr2)
Segregation Distorter	Segregation distorter (Sd)
Zeste-white 3	shaggy (sgg)
Zeste-White 3	shaggy (sgg)
zeste-white 3	shaggy (sgg)

Nous remarquons dans le tableau 25, que des variantes actuellement imprévues sont la conséquence de transformations systématiquement opérées, bien que non prise en charge par le système. Le système pourrait donc être amélioré pour prendre en charge ces transformations.

La présence des labels *Abdominal- A, alpha- spectrin*, *D- Mek* et *lethal-of- scute* s'explique par un formatage particulier des textes de *Medline*. Il s'agit assez vraisemblablement d'un traitement automatique effectué sur les textes qui a pour but de permettre des recherches en texte intégral sur des termes qui participent à des mots composés. Cependant cet espace après le tiret n'est pas systématique.

Nous remarquons aussi à l'œuvre des transformations de concaténation dans les labels *acetylcholinesterase*, *Acetylcholinesterases* et *extramacrochaete*.

Symétriquement les labels *double sex* et *extra sex combs* résultent d'une transformation de scission.

Des opérations de mise en majuscule des premières lettres de certains des mots qui entrent dans la composition d'un label sont aussi à l'œuvre dans *Lethal of Scute* et *Segregation Distorter*.

d. Importance relative des variantes prévues et imprévues

Il est intéressant de comparer l'importance relative des transformations prévues et imprévues. Le tableau ci-dessous donne cette information. Ainsi, en termes d'occurrence, 70 % des définitions variantes sont déjà prise en charge par notre système. Des progrès sont néanmoins encore possible.

Tableau 26 Importance relative des variantes prévues et imprévues
L'effectif correspond au nombre de reconnaissance dans l'échantillon A.

Type	Effectif	Proportion
Prévue	100	70 %
Imprévue	42	30 %
Total	142	100 %

Il est à noter que le nombre de reconnaissances correspondant aux variantes prévues vaut 100, alors que la somme des fréquences mentionnées dans le tableau 24 vaut 118. Cette différence provient du fait qu'une même définition peut être issue de plusieurs type de transformation. Ainsi, lors que l'on ajoute les effectifs du tableau 24, on peu compter plusieurs fois une même reconnaissance, de sorte que le résultat est supérieur à l'effectif total.

E. LES ERREURS DU DICTIONNAIRE

1. Les contradictions du dictionnaire

Le dictionnaire que nous avons utilisé comporte parfois des incohérences. Il se peut que deux définitions utilisent le même label, avec une définition de type *nom complet* ou *symbole* et une autre définition de type *nom synonyme*. Nous dirons dans ce cas que nous avons à faire à des définitions ***contradictoires***. Le tableau 27 donne des exemples.

Tableau 27 Contradiction : cas des *noms synonymes*
Des labels de la première colonne sont les *symboles* ou les *noms complets* des gènes de la deuxième colonne, mais ce sont aussi des synonymes des gènes de la troisième colonne. Les *noms synonymes* sont la cause de beaucoup de contradiction.

Label	Gène 1	Gène 2
bxd	(bxd)	Ultrabithorax (Ubx)
cap	capon (cap)	Chromosome-associated protein (Cap)
cap	capon (cap)	Calphotin (Cpn)
cm	carmine (cm)	crumpled (cmp)
da	daughterless (da)	darky (dar)
dl	dorsal (dl)	duplicated legs (dpl)
E(spl)	Enhancer of split (E(spl))	E(spl) region transcript mbeta (HLHmbeta)
eag	ether a go-go (eag)	eagle (eg)
H1	haemolymph protein 1 (H1)	Histone H1 (His1)
H1	haemolymph protein 1 (H1)	(Su(osk)H1)
Met	Metatarsi irregular (Met)	Resistance to Juvenile Hormone (Rst(1)JH)
mis	misproportioned (mis)	canoe (cno)
PKA	Protein Kinase A (PKA)	(Pka-R1)
PKA	Protein Kinase A (PKA)	cAMP-dependent protein kinase 1 (Pka-C1)
ras	raspberry (ras)	Ras oncogene at 85D (Ras85D)
ras	raspberry (ras)	Ras oncogene at 64B (Ras64B)
shv	shiva (shv)	decapentaplegic (dpp)
shv	shiva (shv)	shortened veins (svs)
Ste	Stellate (Ste)	Suppressor of Stellate (Su(Ste))
zip	zipper (zip)	unzipped (uzip)

Seuls sont présentés dans ce tableau des cas qui ont été rencontrés dans les textes. Cela signifie que tous les labels de la première colonne se trouvent au moins une fois dans les textes que nous avons analysés. Il ne s'agit donc pas a priori de contradictions portant sur les labels extrêmement rares qui ne se rencontrerait pas dans la pratique.

En revanche, le tableau 76 fournis l'ensemble des contradictions du dictionnaire indépendamment du fait qu'elles ont ou n'ont pas étés illustrées dans au moins un texte.

Il ne s'agit pas, à proprement parler, d'erreurs du dictionnaire. Il s'agit seulement d'incohérences dans l'usage, qui ont été fidèlement consignées par *Flybase*. Par exemple, il n'est pas impossible qu'un auteur isolé ait écrit dans un de ses résumés *dl* pour faire référence au gène *duplicated legs (dpl)* alors qu'il est clair que *dl* est déjà pris pour *dorsal (dl)*. Il n'en reste pas moins que nous devons faire la chasse à ces incohérences et empêcher qu'elles nuisent au processus automatique d'identification des gènes dans les textes.

Nous proposons pour cela de mettre de coté les définitions de type *synonyme* qui seraient en conflit avec les définitions de type *nom complet* ou *symbole*. Cette règle d'abord inspirée par le bon sens a été validée comme suit. Nous avons calculé le nombre de fois où cette règle est pertinente dans les textes que nous avons annotés. Cette statistique est faite sur les 112 résumés de l'échantillon A. Elle s'est avéré pertinente dans 54 cas sur 55. La seule exception est donnée dans l'exemple en page 74.

Exemple 10 Préférence donnée à un synonyme
Ce cas de figure est exceptionnel. C'est *Ras oncogene at 85D (Ras85D)* et non *raspberry (ras)* dont il est question dans la dernière phrase. En effet, dans la phrase qui précède, l'auteur fait référence à *Ras oncogene at 85D (Ras85D)* quand il donne un synonyme : *Ras1*.

The expression of the D. melanogaster transcription factor Jun in the eye imaginal disc correlates temporally and spatially with the determination of neuronal photoreceptor fate. Expression of dominant negative forms of Jun in photoreceptor precursor cells results in dose-dependent loss of photoreceptors in the adult fly. Conversely, localized overexpression of Jun in the eye imaginal disc can induce the differentiation of additional photoreceptor cells. Furthermore, the transformation of nonneuronal cone cells into R7 neurons elicited by constitutively active forms of sevenless, Ras1, Raf, and MAP kinase is relieved in the presence of Jun mutants. These results demonstrate a requirement of Jun downstream of the sevenless/ras signaling pathway for neuronal development in the Drosophila eye.

Nous venons de voir les cas de contradictions dans le dictionnaire qui ont lieu entre des définitions de type *nom complet* et des définitions de type *nom synonyme*. Il existe aussi, mais ils sont plus rares, des cas de conflits entre définitions de type *symbole* et de type *nom complet*. Le tableau 28 fournit les deux cas qui se trouvent dans le dictionnaire.

Tableau 28 Contradiction entre symbole et *nom complet*.
Le label de la première colonne est le symbole du gène de la seconde colonne, mais aussi le *nom complet* du gène de la troisième colonne. Seul deux cas sont présents dans le dictionnaire issu de *Flybase*.

Label	Gène 1	Gène 2
Apc	APC-like (Apc)	Apc (Fs(3)Apc)
dark	darkener of white-eosin (dark)	dark (dk)

2. Des définitions aberrantes

Nous avons vu que tout emploi, même marginal, est consigné dans le dictionnaire *Flybase*. Ces informations sont parfois trop anecdotiques pour constituer une bonne définition de gène. Par exemple, il est noté dans *Flybase* que le terme *transcript group V*, et même la lettre V (qui représente le chiffre romain cinq) a été utilisé par un auteur dans un texte pour désigner le gène *engrailed*. Nous ne pouvons cependant pas en conclure que nous sommes en présence du gène *engrailed* à chaque fois que nous rencontrons dans un texte le terme *transcript group* ou la lettre V. Ce type d'information constitue ce que nous appellerons une définition ***aberrante***. Le tableau 37 donne quelques exemples de telles définitions.

Le problème pour nous, réside dans le fait que de telles définitions, inutiles et même gênantes, sont mêlées à des définitions pertinentes. En effet, tout ce que l'on peut dire est qu'elles sont classées dans la catégorie des définitions de type *nom synonyme*. Or il y a aussi dans cette catégorie de bonnes définitions comme par exemple celles qui correspondent à des variations orthographiques sur le *symbole* ou le nom complet d'un gène.

3. Les formats imprévus

Pour finir sur les problèmes rencontrés dans le dictionnaire issu de *Flybase*, nous signalons des erreurs ou irrégularité dans le format. Par exemple, à l'intérieur du champ *nom synonyme* on trouve des explications complémentaires sur le label présenté. Il peut s'agir d'un commentaire placé entre parenthèse qui indique que le nom de gène est déjà utilisé dans une autre définition ou qu'il est suspect. On trouve aussi des explications sur l'origine des labels qui consistent à donner pour un *nom synonyme* le label dont il est l'abréviation. Des références bibliographiques sommaires sont aussi parfois données entre parenthèses. On

trouve aussi le terme *unnamed*, voire le point d'interrogation. Tout ceci n'est pas systématique et peut difficilement être exploité. Il est donc nécessaire de filtrer cette information.

Nous signalons aussi des problèmes de doublons, c'est à dire de redondance de l'information. On trouve ainsi assez fréquemment des définitions qui sont répétées plusieurs fois pour un même gène avec, par exemple, une des définitions de type *nom complet* et l'autre définition de type *nom synonyme*. Ces informations inutiles, car redondantes, ont été supprimées de façon à faciliter les opérations de dénombrement.

Dans le même ordre d'idée, nous avons aussi eu quelque cas de doublons. Par exemple, de gène *popeye (pop)* (FBgn 14375[30]) est identique à *popeye* (FBgn 3125). Nous avons donc mis de côté le second gène.

F. NÉCESSITÉ DE L'UTILISATION DU CONTEXTE

1. Utilisation du contexte pour préférer une reconnaissance à une autre

Dans certain cas, nous voyons que c'est le contexte qui permet de comprendre le sens du texte. Voici un exemple ci-dessous.

Exemple 11 Interprétation et contexte
Le contexte est nécessaire pour l'interprétation. Ici le terme souligné *Nos* désigne non pas le gène *Nitric oxide synthase (Nos)*, mais la protéine du gène *nanos (nos)*.

Localization of the maternally synthesized nanos (nos) RNA to the posterior pole of the Drosophila embryo provides the source for a posterior-to- anterior gradient of Nos protein. Correct spatial regulation of nos activity is essential for normal pattern formation. High local concentrations of Nos protein in the posterior of the embryo are necessary to inhibit translation of the transcription factor Hunchback in this region, and thus permit expression of genes required for abdomen formation (see ref. 5 for review). By contrast, misexpression of Nos protein at the anterior of the embryo prevents translation of the anterior morphogen Bicoid, suppressing head and thorax development. Posterior localization of nos RNA is mediated by sequences within the nos 3' untranslated region (3'UTR) and requires the function of eight genes of the 'posterior group'. Although the unlocalized nos RNA is stable in embryos from females mutant for any of the posterior group genes, these embryos appear to lack nos activity because they develop the abdominal defects characteristic of embryos produced by nos mutant females. We report here that unlocalized nos RNA is translationally repressed. Translational repression is mediated by the nos 3'UTR and can be alleviated either by replacement of the 3'UTR with heterologous 3'UTR sequences or by posterior localization. Thus, RNA localization provides a novel mechanism for translational regulation.

Dans cet exemple, nous voyons que paradoxalement une définition de type *nom synonyme* (plus précisément du type *protéine*) doit être préférée à une définition de type *symbole*. Nous constatons que cette utilisation du contexte pour interpréter le label doit être faite au niveau du résumé entier et pas au niveau de la phrase.

Dans le précédent exemple, nous avons vu une préférence donnée à un gène sur un autre gène dans l'interprétation d'un label. Il se peut aussi que l'on donne, grâce au contexte, la préférence à un *complexe de protéines* au détriment d'un *gène*. Ceci est illustré dans l'exemple ci-après.

[30] Flybase donne un numéro unique à chacun des gènes qu'il répertorie. Son format est du type FBgnxxx.

Exemple 12 Utilisation du contexte : cas d'un complexe de protéine
Un *complexe de protéines* est préféré à un gène par l'utilisation du contexte. Ici les occurrences soulignées de *PS* doivent être interprétées comme des références au complexe *PS integrins (PS)* et non au gène *Presenilin (PS)*.

The two Drosophila position-specific (PS) integrins are expressed on complementary sides of sites where different cell layers adhere to each other, such as the attachments of the embryonic muscles to the epidermis. While there is suggestive evidence that the PS integrin-mediated adhesion is via the extracellular matrix, it is also possible that it occurs through the direct interaction of the two integrins, alpha PS1 beta PS and alpha PS2 beta PS. To help distinguish between these possibilities a comparison between the phenotypes caused by the absence of the beta PS subunit and the absence of one of the PS alpha subunits, alpha PS2, has been made. Two pieces of evidence are provided that prove that the alpha PS2 subunit is encoded by the locus inflated (if). Firstly, three new if alleles have been isolated, each of which is associated with a molecular lesion in the alpha PS2 gene, and each of which results in the complete loss of if activity. Secondly, a 39 kb fragment of genomic DNA that encompasses the alpha PS2 gene completely rescues if mutations when introduced into the germline by P- element-mediated transformation. A comparison of the null inflated phenotype with that of the locus that encodes the beta PS subunit, myospheroid (mys), reveals that while the beta PS subunit is required for the adhesion of the epidermis along the dorsal midline, the alpha PS2 subunit is not. In if mutant embryos, the muscles remain attached to the other cell layers significantly longer than in a mys mutant embryo. This shows that the alpha PS2 beta PS integrin only contributes part of the adhesive activity at the sites of PS integrin adhesion, and rules out a model where PS integrin function occurs solely by the direct interaction of the two PS integrins.

Le contexte permet aussi de lever l'ambiguïté qu'il peut y avoir avec des allèles. Voici un exemple ci-dessous.

Exemple 13 Utilisation du contexte : cas des Allèles
Un allèle est préféré à un gène par l'utilisation du contexte. Les termes soulignés *wbl* et *wa* font référence aux allèles *white-blood (wbl)* et *white-apricot (wa)* du gène *white (w)* et non aux gènes *windbeutel (wbl)* et *warty (wa)*.

We are interested in identifying single gene mutations that are involved in trans-acting dosage regulation in order to understand further the role of such genes in aneuploid syndromes, various types of dosage compensation as well as in regulatory mechanisms. The Lighten up (Lip) gene in Drosophila melanogaster was identified in a mutagenic screen to detect dominant second site modifiers of white-blood (wbl), a retrotransposon induced allele of the white eye color locus. Lip specifically enhances the phenotype of wbl as well as a subset of other retroelement insertion alleles of white, including the copia-induced allele, white-apricot (wa), and six alleles caused by insertion of I elements. We isolated six alleles of Lip which are all recessive lethal, although phenotypically additive heteroallelic escapers were recovered in some combinations. Lip also suppresses position effect variegation, indicating that it may have a role in chromatin configuration. Additionally, Lip modifies the total transcript abundance of both the blood and copia retrotransposons, having an inverse effect on the steady state level of blood transcripts, while showing a non-additive effect on copia RNA.

Nous constatons dans cet exemple que, bien que la terminologie des gènes de la drosophile ait été faite avec suffisamment de rigueur pour qu'un *symbole* ne renvoie toujours qu'à un gène, cela n'est pas suffisant, car quand on regarde les choses au niveau des gènes et des allèles, il y a des ambiguïtés. Ici par exemple l'allèle *white-abricot (wa)* a un *symbole* qui se confond avec le *symbole* de *warty (wa)*.

Dans un dernier cas de figure, ce que l'on préfèrera à un gène n'a rien à voir avec un gène. C'est ce que nous appellerons un ***objet spécifique***. Le voici illustré en page 77.

Exemple 14 Utilisation du contexte : cas d'un objet spécifique
Un objet spécifique est préféré à un gène par l'utilisation du contexte. Le terme souligné *AR* est l'abréviation de *adaptative response*, il ne doit pas être interprété comme une référence au gène *Adrenodoxin reductase (AR)*.

The effects of a low dose (0.1-20 mGy) preirradiation with X-rays followed by a higher dose (2 Gy) of the same radiation on the recovery of the genetic damage induced as dominant lethals in mature oocytes (stage 14) of different strains of Drosophila melanogaster were investigated. The response was shown to be dependent on the genotype of the flies tested, since lower frequencies of dominant lethals (DL) were only obtained in strains carrying the white mutation. Based on these observations experiments to locate the genetic factor responsible for the adaptive response (AR) were performed. This factor was found to be in a specific region of the X-chromosome. Additional experiments were carried out to give information on the minimal dose required to induce the AR. The results showed that the lowest dose needed is 0.2 mGy. Increasing the conditioning X-ray dose had no influence on the response.

Nous voyons dans cet exemple que l'objet spécifique *adaptive response (AR)* se comporte exactement comme un gène. Il sera donc intégré au dictionnaire comme les gènes, les complexes de protéines, les complexes de gènes et famille de gènes et les allèles. Cependant, une rubrique spéciale lui sera réservée.

L'étude de ces exemples nous permet de donner quelques définitions et d'établir des règles pour la reconnaissance des définitions.

Quand la reconnaissance d'une définition dans un texte est crédibilisée par la reconnaissance d'une autre définition du même gène, nous dirons que la reconnaissance est ***confirmée***. Dans le cas contraire nous dirons qu'elle est ***isolée***.

Dans le cas où, à une occurrence donnée d'un label dans un texte, n'est associée qu'une seule reconnaissance, nous dirons que cette reconnaissance est ***simple***. Dans le cas contraire nous dirons qu'elle est ***multiple***.

Nous avons vu que dans le cas de certaines contradictions du dictionnaire, il est nécessaire de privilégier une définition sur une autre. Par exemple, le label *Nos* sera préférentiellement associé au gène *Nitric oxide synthase (Nos)* plutôt qu'au gène *nanos (nos)*. La définition A, qui associe *Nos* à *Nitric oxide synthase (Nos)* sera caractérisée de ***privilégiée***, ce qui signifie qu'elle sera en général préférée à la définition B, qui associe *Nos* à *nanos (nos)*. Il existe cependant des cas comme dans l'exemple en page 75, où la reconnaissance de la définition B est confirmée, alors que la reconnaissance de la définition A ne l'est pas. Dans ce cas, il faut effectuer la reconnaissance de B et ne pas effectuer celle de A.

Nous dirons donc que la définition B est ***à confirmer***. Cela signifie qu'elle ne sera reconnue que s'il y a confirmation.

Nous voyons dans cet exemple que la reconnaissance de A ne se fait pas, parce qu'elle est ***multiple*** et qu'elle est ***non confirmée***.

Nous venons de voir que le contexte permet de résoudre les problèmes des contradictions qui se trouvent dans le dictionnaire. Voyons maintenant comment nous pouvons l'utiliser pour résoudre le problème des labels ambigus.

2. Utilisation du contexte pour régler le problème de l'ambiguïté des labels

Nous avons vu que certains labels présentent un caractère ambigu, à savoir qu'ils ne renvoient pas nécessairement vers un gène. Voyons sur l'exemple en page 78 comment le contexte permet de résoudre le problème.

Exemple 15 Contexte et ambiguïté des labels
La confirmation de la reconnaissance d'une définition permet de lever l'ambiguïté sur le label. Ici l'ambiguïté sur le label *stripe* (souligné) est levée par la présence du label *sr* (souligné), qui participe à la définition du même gène, à savoir *stripe (sr)*.

The different thoracic muscles of Drosophila are affected specifically in the mutants: stripe (sr), erect wing (ewg), vertical wings (vtw), and nonjumper (nj). We have tested the extent of this specificity by means of a genetic analysis of these loci, multiple mutant combinations, and gene dosage experiments. A quantitative, rather than a qualitative, specificity is found in the mutant phenotypes. All muscles are altered by mutations in any given gene, but the severity of these alterations is muscle specific. The locus stripe seems to have a polar organization where different allelic combinations show quantitative specificity in the muscle affected. In addition to the muscle phenotypes, neural alterations are detected in these mutants. The synergism found between ewg, vtw and ewg, sr as well as the dosage effect of the distal end of the X chromosome upon the expression of ewg and sr suggests the existence of functional relationships among the loci analyzed.

Certains labels renvoient trop souvent vers autre chose qu'un gène pour permettre, à eux seul, de reconnaître un gène. En revanche, dés lors que la reconnaissance de la définition à laquelle ils participent est confirmée, on peut les interpréter. Nous dirons que ces labels sont ***à confirmer***.

Le même procédé peut être utilisé pour les *mots vides* comme illustré dans l'exemple ci-dessous.

Exemple 16 Utilisation du contexte : cas des *mots vides*
Même les *mots vides* peuvent être interprétés quand ils sont confirmés. Ici *if* renvoie bien à *inflated (if)* comme en témoigne la présence dans le texte des labels *inflated* et *alpha PS2*.

The two Drosophila position-specific (PS) integrins are expressed on complementary sides of sites where different cell layers adhere to each other, such as the attachments of the embryonic muscles to the epidermis. While there is suggestive evidence that the PS integrin-mediated adhesion is via the extracellular matrix, it is also possible that it occurs through the direct interaction of the two integrins, alpha PS1 beta PS and alpha PS2 beta PS. To help distinguish between these possibilities a comparison between the phenotypes caused by the absence of the beta PS subunit and the absence of one of the PS alpha subunits, alpha PS2, has been made. Two pieces of evidence are provided that prove that the alpha PS2 subunit is encoded by the locus inflated (if). Firstly, three new if alleles have been isolated, each of which is associated with a molecular lesion in the alpha PS2 gene, and each of which results in the complete loss of if activity. Secondly, a 39 kb fragment of genomic DNA that encompasses the alpha PS2 gene completely rescues if mutations when introduced into the germline by P- element-mediated transformation. A comparison of the null inflated phenotype with that of the locus that encodes the beta PS subunit, myospheroid (mys), reveals that while the beta PS subunit is required for the adhesion of the epidermis along the dorsal midline, the alpha PS2 subunit is not. In if mutant embryos, the muscles remain attached to the other cell layers significantly longer than in a mys mutant embryo. This shows that the alpha PS2 beta PS integrin only contributes part of the adhesive activity at the sites of PS integrin adhesion, and rules out a model where PS integrin function occurs solely by the direct interaction of the two PS integrins.

Le risque de faire une erreur en interprétant à tort un ***mot vide*** est faible car les définitions relatives aux *mots vides* ont été examinées et trois d'entres-elles ont été invalidées comme illustré dans le tableau 29.

Tableau 29 Mots vides : définitions invalidées
Les définitions relatives aux *mots vides* et qui ne sont pas crédibles ont été désactivées.

Label	Gène
be	tumor(3)be (tu(3)be)
do	pale ocelli (po)
in	inturned (in)

3. Utilisation du contexte pour détecter les reconnaissances redondantes

Il est fréquent de voir un auteur apporter des précisions sur la terminologie en donnant pour le même gène, dans la même phrase et l'un à la suite de l'autre, deux de ces labels. Typiquement, la première occurrence donne le *nom complet* et la seconde le *symbole*, ce dernier étant placé entre parenthèses. Dans ce cas, nous dirons que la deuxième reconnaissance est ***redondante***. Le voici illustré dans l'exemple ci-dessous.

Exemple 17 Les reconnaissances redondantes.
Le gène *decapentaplegic (dpp)* est reconnu deux fois dans cette phrase, une première fois par le label *decapentaplegic*, puis une deuxième fois par le label *dpp*. La deuxième reconnaissance est dite redondante car elle suit immédiatement la première.

The decapentaplegic (dpp) gene in Drosophila melanogaster encodes a TGF- beta-like signalling molecule that is expressed in a complex and changing pattern during development.

Il est important de savoir repérer ce type de reconnaissance pour l'extraction d'information sur les interactions génétiques. En effet, il ne faudrait pas considérer les deux occurrences consécutives comme des partenaires d'une éventuelle interaction. Les reconnaissances redondantes ne seront pas prises en compte dans la recherche de partenaires.

La redondance est un phénomène très fréquent. Dans l'échantillon A, 109 reconnaissances sur 1417 (soit 8 %) sont redondantes. Cela correspond à 62 résumés soit 52 % des 112 résumés que compte l'échantillon A.

4. Utilisation du contexte pour valider les définitions CRÉÉS pour anticiper les variations orthographiques des labels

Nous avons vu que le dictionnaire des gènes n'est pas complet. Nous l'avons complété automatiquement en ajoutant des définitions qui sont des variantes des définitions originales. Ce processus peut malheureusement conduire à créer de nouvelles difficultés en introduisant des labels ambigus c'est à dire qui désignent éventuellement autre chose qu'un gène dans les textes que nous analysons. Le tableau 30 donne des exemples de définitions qui ne sont pas correctes, car les labels sont ambigus.

Tableau 30 Invalidation des variantes non confirmés
Les définitions créées automatiquement et qui ne sont confirmés dans aucun des textes sont présentées ici. La dernière colonne donne le nombre d'occurrence du label dans les textes. Le tableau complet, obtenu par l'analyse automatique de 744 résumés issus de Medline compte 137 lignes. Il est clair que ces définitions ne doivent pas être prises en compte.

Label	Gène	Fréquence
to	Superoxide dismutase (Sod)	205
is	Isis (Is)	177
D	dachs (d)	69
C	curved (c)	62
on	Open (On)	37
large	Large (Lg)	34
bristle	Bristle (Bl)	23
set	Set	18
AS	ascute (as)	17
T	tan (t)	17
margin	Margin (Mar)	14
G	garnet (g)	12
viability	Ribosomal protein L36 (RpL36)	11
lethals	LETHALS	10
open	Open (On)	10
P	pink (p)	10

Il est nécessaire de valider, d'une façon ou d'une autre, les définitions que nous avons rajoutées au dictionnaire. Nous avons choisi de désactiver les définitions qui ne sont confirmées dans aucun des 744 résumés que nous avons analysés.

II. ANALYSE DU PROBLÈME DE LA RECONNAISSANCE DES INTERACTIONS

Le principe de l'analyse a été posé dans la section Partie 1Chapitre 3I.C.2 et notamment dans l'équation 2. Il nous reste à préciser la méthode, et notamment dire comment d'une part prendre en compte la présence d'un ou plusieurs gènes dans une phrase et d'autre part, comment combiner cette information avec l'existence d'un *IVI* favorable pour arriver à extraire l'information.

Nous analyserons tout d'abord (partie A) la question de l'extraction d'information sur les interactions, indépendamment de la méthode mise en œuvre ; avant de passer (partie B) aux difficultés que pose la méthode que nous proposons.

A. COMPLEXITÉ DE LA RECONNAISSANCE DES INTERACTIONS

Cette complexité tient à la nature même de la tâche, indépendamment de la méthode mise en œuvre pour y parvenir.

Il s'agit en particulier de trouver une définition pertinente de la notion d'interaction. En effet, dans le travail de PILLET, on est parti du principe que les interactions étaient représentées par des couples de gènes. Cela découlait en réalité d'une technique d'extraction de l'information qui voulait que l'on ne s'intéresse qu'aux seules phrases qui contiennent exactement deux occurrences de gènes. Le travail sur *Medline* s'est fait sans ce présupposé. C'est à dire que l'on a annoté toutes les phrases, quel que soit le nombre de gènes cités dans celles-ci. On se rend compte alors, à la lecture des textes, qu'il est parfois difficile de réduire l'information contenue dans une phrase à des listes de couples de gènes en interaction.

Il est vrai que l'on peut concevoir un ensemble de gènes en interaction comme un réseau de gènes en interactions moléculaires. Cependant, certains des chaînons de ce réseau font parfois intervenir plus de deux gènes. On ne peut donc réduire le réseau à une liste d'interaction deux à deux. De plus, les faits décrits dans les textes ne portent pas toujours sur des interactions moléculaires. Plus précisément, il s'agit souvent de chemins de signalisation et de groupes de gènes en interaction, ce qui ne peut se ramener entièrement à un ensemble d'interactions direct ou non.

1. Partenaires mal définis

Il arrive fréquemment que les auteurs énoncent des interactions entre des objets qui ne sont pas des gènes à proprement parler, mais des *collections* de gènes, c'est à dire des objets de type famille de protéines, complexe de protéines ou complexe de gènes. L'exemple ci-dessous illustre le cas particulier d'une interaction faisant intervenir un complexe de gènes.

Exemple 18 Interaction faisant intervenir des groupes de gènes
Cette phrase énonce une interaction entre le complexe de gènes *Achaete-scute Complex (ASC)* et le gène *daughterless (da).*

The products of the achaete-scute complex and daughterless interact to form heterodimers able to activate transcription

Cette information ne peut, en toute rigueur, se traduire par l'ensemble des interactions entre le gène *daughterless* et les membres du complexe *Achaete-scute*. Cependant le vocabulaire de l'interaction est bien présent avec ce type d'énoncé, de sorte que l'on risque d'introduire une distorsion dans les statistiques si on ne note aucune interaction.

La solution idéale consiste à mentionner l'interaction même si elle fait intervenir un groupe de gènes comme un complexe. C'est aussi l'intérêt d'un système de reconnaissance des gènes qui prennent en compte les complexes de gènes. Une solution alternative, et qui a beaucoup été employée lors de l'annotation manuelle, consiste à noter une interaction mais sans préciser tous les partenaires de celle-ci. Dans le cas présent seul le partenaire *daughterless* sera noté, l'autre partenaire restant vide.

L'exemple suivant illustre le cas d'une interaction entre un gène d'une part, et une famille de protéines d'autre part. Cette information ne peut pas se traduire en toute rigueur par un ensemble d'interactions entre d'une part, le gène et d'autre part, les gènes qui codent pour chacun des éléments de la famille de protéines.

Exemple 19 Interaction faisant intervenir des familles de protéines
Cette phrase énonce une interaction entre le gène *TATA binding protein (Tbp)* et la famille de protéines *Heat-shock-protein-70 (hsp70).*

Immunopurified TFIID produces a large DNase I footprint over the hsp70, hsp26, and histone H3 promoters of Drosophila

Ce type d'information n'a pas été systématiquement pris en compte par les annotateurs. En conséquence on ne connaît pas l'importance numérique de ce type d'information.

2. Interaction et ordre

Dans certains cas, le texte ne précise pas le sens de l'interaction. Ceci est illustré dans l'exemple en page 82.

Exemple 20 Interaction non ordonnée
Dans cette phrase, le sens de l'interaction entre le gène *brahma (brm)* et le gène *absent, small, or homeotic discs 1 (ash1)* n'est pas précisé.

Mutations in the gene brahma (brm) which also is one of the trithorax set of genes interact with mutations in ash1 such that non-lethal ash1 +/+ brm double heterozygotes have a high penetrance of homeotic transformations in specific imaginal disc- and histoblast-derived tissues

Les interactions seront donc caractérisées d'***ordonnées*** ou de ***non ordonnées***. Quand il est possible de dire à la lecture du texte quel est l'ordre, c'est à dire de préciser quel est le gène qui a une influence sur l'expression de l'autre gène, alors cette information est consignée dans une interaction ordonnée. Dans la base coexistent des interactions ordonnées et des interactions non ordonnées. Cependant nous n'avons pas cherché, lors de l'annotation automatique, à proposer un ordre. Autrement dit, les annotations faites par le programme sont toutes de type non-ordonnées. La confrontation des annotations de l'expert avec celles de la machine reste cependant possible car toute interaction ordonnée peut être dégradée en une interaction non ordonnée : il suffit de conserver l'information sur les partenaires et d'oublier celle qui concerne l'ordre. Nous verrons mieux cela dans la partie résolution des problèmes dans la section Chapitre 2II.A.

Dans les textes, le sens des interactions est généralement précisé, comme on peut le constater au vu des chiffres du tableau 31.

Tableau 31 Interaction et ordre
Chaque interaction est caractérisée d'ordonnée ou non. Les interactions sans précision d'ordre sont moins fréquentes mais ne sont pas quantité négligeable. Les chiffres sont issus de l'annotation manuelle de l'échantillon A.

Ordre précisé	Fréquence	Proportion
Oui	310	92%
Non	28	8%
Total	*338*	*100%*

L'ordre a été précisé à chaque fois que cela était possible dans l'annotation des textes. D'autres informations auraient pu l'être mais ne l'ont pas été, telles que le sens (activation ou inhibition), le caractère moléculaire ou génétique de l'interaction ou le type d'expérience ayant mis en évidence l'interaction.

3. Partenaires de l'interaction non identifiés

Dans certains cas, l'auteur ne précise pas quels sont les gènes en présence dans une interaction. Cela est illustré dans l'exemple suivant.

Exemple 21 Partenaires de l'interaction non identifiés
Dans cette phrase, des interactions entre vasa et quatre autres gènes sont énoncées, mais on ne dit pas quels sont les partenaires en question.

We have found that localization of vasa to the perinuclear nuage is abolished in most vas alleles, but is unaffected by mutations in four genes required upstream for its pole plasm localization

La question est de savoir s'il faut prendre en compte ou non ce type d'information, et si oui comment. Il serait préjudiciable de ne pas du tout en tenir compte. En effet, ce type de phrase, même si elle n'énonce pas une interaction claire, utilise quand même le vocabulaire de l'interaction. Ainsi, lors de l'étude du vocabulaire de l'interaction, il faut considérer ce type de phrase comme un énoncé d'interaction.

B. DIFFICULTÉ DE LA RECONNAISSANCE DES INTERACTIONS

Dans la section A ci-dessus, nous avons axé notre réflexion sur la difficulté de l'extraction d'informations indépendamment de la méthode mise en œuvre pour extraire cette information. Dans la présente section, nous axons notre réflexion sur les difficultés qui peuvent survenir à cause de la méthode mise en œuvre.

1. Partenaires de l'interaction absents de la phrase mais pas du résumé

L'analyse des textes se fait par segments de texte. Nous avons choisi d'annoter les textes au niveau des phrases, sachant que grâce au schéma relationnel il reste possible de consulter les annotations par résumé. Il parait naturel de vouloir extraire l'information en considérant indépendamment les phrases d'un même résumé. Dans ce cas, on est confronté à l'existence de phrases, qui ne citent pas les partenaires des interactions qu'elles décrivent. Ceci est illustré dans l'exemple suivant.

Exemple 22 Partenaires non-cités dans la phrase
L'occurrence soulignée du terme *allèles*, fait référence aux allèles du gène *tolloid (tld)* mais cela n'est précisé que dans la phrase qui précède. Si l'analyse se fait phrase par phrase, l'interaction énoncée entre *tolloid (tld)* et *decapentaplegic (dpp)* ne pourra pas être détectée.

To characterize the nature of the tolloid mutations, we have sequenced eighteen tolloid alleles. We find that five of the seven alleles that act as dominant enhancers of dpp are missense mutations in the protease domain.

Le phénomène est assez rare puisqu'il n'a été constaté par l'annotateur que pour 8 interactions parmi les 2766 faites sur l'échantillon A.

2. Difficulté introduite par la présence de plus de deux gènes dans une même phrase

Nous avons constaté l'existence de descriptions d'interaction dans des phrases qui comptent plus de deux occurrences de gène. Ceci est illustré dans l'exemple suivant.

Exemple 23 Plus de deux gènes dans une même phrase.
Dans cette phrase, 9 gènes sont cités et il est question d'interaction mais elles ne sont qu'au nombre de 9, alors qu'il pourrait y en avoir jusqu'à 72 si chaque gène en présence interagissait avec chacun des autres.

Males hemizygous for a temperature-sensitive allele, ph2, are lethal when heterozygous with mutants in Asx, Pc, Pcl, Psc, Sce and Scm, and with E(Pc) and Su(z)2.

Dans les phrases qui ne comptent que deux occurrences de gènes, la situation est simple : s'il y a interaction, on connaît immédiatement les deux partenaires. Des exceptions existent car un des partenaires d'une interaction peut ne pas être cité, mais elles sont rares. En revanche, dans le cas d'une phrase pour laquelle il existe d'avantages d'occurrences de gènes, on ne peut pas prévoir entre quels gènes il y a interaction, à supposer qu'il y en ait.

Ce type de problème est important car comme on peut le voir dans le tableau 32, seulement 31% des reconnaissances d'interactions sont issues d'une phrase qui contient exactement deux occurrences de gène.

Tableau 32 Interaction et nombre d'occurrence de gène

Le tableau donne le nombre de reconnaissance d'interactions (colonne du milieu) en fonction du nombre d'occurrence de gènes dans la phrase (colonne de gauche). Cette statistique est faite à partir des annotations manuelles de l'échantillon A.

Gène	Interaction	Proportion
0	0	0%
1	13	5%
2	75	31%
3	53	22%
4	43	18%
5	35	15%
6 ou plus	22	9%
0 ou plus	*241*	*100%*

Tableau 33 Labels assez ambigus
Les labels présentés prêtent à confusion avec des termes d'anglais courant. Le gène désigné est donné dans la colonne de droite.

Label	Gène
abdominal	abdominal (abd)
act	actidione-sensitive (act)
al	aristaless (al)
ambiguous	ambiguous (abg)
antenna	empty spiracles (ems)
arrest	arrest (aret)
band	band (bn)
blocked	blocked (blc)
blood	blood element (blood)
bp	bulge (bul)
brief	brief (bf)
broad	broad (br)
c	curved (c)
condensed	condensed (con)
cortex	cortex (cort)
dark	dark (dk)
dark	darkener of white-eosin (dark)
defective	defective (df)
depleted	depleted (ded)
depressed	depressed (dep)
divergent	divergent (dv)
double	(double)
early	early (eay)
early	lodestar (lds)
ectodermal	ectodermal (ect)
extended	extended (ext)
eye	clift (cli)
eye	eyeless (ey)
h	hairy (h)
high	high (hi)
inactive	inactive (iav)
juvenile	juvenile (juv)
labial	labial (lab)
leg	runt (run)
limited	limited (lm)
lines	lines (lin)
ll	lanceolate (ll)
M	(anon-ESTCL2c12)
map	midgut amylase pattern PMG (mapP)
mid	midline (mid)
midline	midline (mid)
midway	midway (mdy)
missing	missing (msg)
multiple	multiple (mul)
N	Notch (N)
naked cuticle	naked cuticle (nkd)
narrow	narrow (nw)
paired	paired (prd)
period	period (per)
pre	presto (pre)
R	Roughened (R)
r	rudimentary (r)
raised	raised (rsd)
ray	rayon (ray)
re	reduced eyes (re)
re	rough eye (rey)
reduced	reduced (rd)
retained	retained (retn)
reversed polarity	reversed polarity (repo)
rudimentary	rudimentary (r)
s	sable (s)
SD	Segregation distorter (Sd)
separated	separated (sep)
shifted	shifted (shf)
similar	similar (sima)
small	small (sml)
smaller	smaller (sma)
spliced	torso (tor)
spread	spread (sprd)
stripe	stripe (sr)
stripes	stripes (str)
terminus	terminus (term)
trunk	trunk (trk)
twisted	twisted (tw)
ve	rhomboid (rho)
vein	vein (vn)

Tableau 34 Labels faiblement ambigus.
Dans le doute, ils serviront à l'identification des gènes dans les textes.

Label	Gène	Remarque
abbreviated	abbreviated (abb)	
abrupt	abrupt (ab)	
adipose	adipose (adp)	
amalgam	Amalgam (Ama)	
Amalgam	Amalgam (Ama)	
amber	amber (amb)	signifie ambre, terme rare
approximated	approximated (app)	
Attenuated	Attenuated (At)	
Bag	Bag (Bg)	
bent	bent (bt)	signifie tordu, terme rare
blunt	blunt (blu)	signifie émoussé, terme rare
bordered	bordered (bord)	signifie en bordure de, terme rare
Bristle	Bristle (Bl)	signifie soie de porc, terme rare
cardinal	cardinal (cd)	
clipped	clipped (cp)	signfie taillé, terme rare
compressed	compressed (com)	
Dense	Dense (Dns)	ne se confond pas avec dense
displaced	displaced (dd)	
Drop	Drop (Dr)	
erratic	erratic (err)	
expanded	expanded (ex)	signifie élargir, terme rare
Large	Large (Lg)	
Malformed	Malformed (Mal)	
miniature	miniature (m)	
Minute	Minute (M(1)7C)	ne se confond pas avec minute
morula	morula (mr)	
oblique	oblique (ob)	
opaque	opaque (op)	
Open	Open (On)	ne se confond pas avec open
pale	pale (ple)	
pointed	pointed (pnt)	
silver	silver (svr)	
Spread	Spread (Spr)	ne se confond pas avec spread
sticky	sticky (sti)	signifie colant, terme rare
Streak	Streak (Sk)	ne se confond pas avec streak
Stripe	stripe (sr)	ne se confond pas avec stripe
syndrome	syndrome (syn)	
thread	thread (th)	signifie fil, terme rare
Tilt	tilt (tt)	signifie inclinaison, terme rare
Tiny	tiny (ty)	signifie foin, terme rare
Tumor	Tumor (TU)	ne se confond pas avec tumor
uncoordinated	uncoordinated (unc)	
uneven	uneven (un)	signifie impair, terme rare
Unfolded	Unfolded (Uf)	ne se confond pas avec unfolded
Vein	Vein (Vn)	
daughterless	daughterless (da)	Signifie sans fille

Tableau 35 Collection de gènes

Les labels présentés désignent une *collection* de gènes. Chacun de ces labels a été trouvé au moins une fois au cours de l'annotation. Chaque ligne du tableau correspond à une définition présente dans le dictionnaire issu de *Flybase*. Ces définitions étant imprécises, elles sont mises de côté dans la reconnaissance des gènes.

Label	Gène
Spectrin	alpha Spectrin (alpha-Spec)
Spectrin	beta Spectrin (beta-Spec)
spectrin	alpha Spectrin (alpha-Spec)
spectrin	beta Spectrin (beta-Spec)
ASC	asense (ase)
ASC	achaete (ac)
BXC	abdominal A (abd-A)
BXC	Abdominal B (Abd-B)
BXC	Ultrabithorax (Ubx)
Actin	Actin 42A (Act42A)
Actin	Actin 57B (Act57B)
Actin	Actin 79B (Act79B)
Actin	Actin 87E (Act87E)
Actin	Actin 88F (Act88F)
Actin	Actin 5C (Act5C)
actin	Actin 42A (Act42A)
actin	Actin 5C (Act5C)
hsp70	Heat-shock-protein-70Ab (Hsp70Ab)
hsp70	Heat-shock-protein-70Ba (Hsp70Ba)
hsp70	Heat-shock-protein-70Bb (Hsp70Bb)
hsp70	Heat-shock-protein-70Aa (Hsp70Aa)
hsp70	Heat-shock-protein-70Bc (Hsp70Bc)
U2AF	(U2af50)
U2AF	(U2af65)
U2AF	(U2af35)
histone	Histone H2A (His2A)
histone	Histone H1 (His1)
histone	Histone H2B (His2B)
histone	Histone H4 (His4)
histone	Histone H3 (His3)
gooseberry	gooseberry distal (gsb-d)
gooseberry	gooseberry proximal (gsb-p)
Amylase	Amylase distal (Amy-d)
Amylase	Amylase proximal (Amy-p)
amylase	Amylase proximal (Amy-p)
amylase	Amylase distal (Amy-d)

Tableau 36 Orthographie absentes de Flybase

Les deux premières colonnes indiquent la définition absente du dictionnaire extrait de *Flybase*, et la troisième le nombre de reconnaissance faite par l'annotateur.

Label	Gène	Fréquence
Wingless	wingless (wg)	12
Armadillo	armadillo (arm)	11
SUHW	suppressor of Hairy wing (su(Hw))	9
Dorsal	dorsal (dl)	8
extramacrochaete	extra macrochaetae (emc)	6
even-skipped	even skipped (eve)	6
UBX	Ultrabithorax (Ubx)	5
dU2AF50	U2af50	5
DPP	decapentaplegic (dpp)	4
calmodulin	Calmodulin (Cam)	4
Bicoid	bicoid (bcd)	3
Abdominal-B	Abdominal B (Abd-B)	3
Tube	tube (tub)	3
Pelle	pelle (pll)	3
Hairless	Hairless (H)	3
Dm cdc2c	cdc2c	3
gooseberry-distal	gooseberry distal (gsb-d)	3
gooseberry-proximal	gooseberry proximal (gsb-p)	3
abdominal-A	abdominal A (abd-A)	3
Nos	nanos (nos)	3
beta 1 tubulin	betaTubulin56D (betaTub56D)	3
cyclin E	Cyclin E (CycE)	3
Sry delta	Serendipity delta (Sry-delta)	3
Sex-lethal	Sex lethal (Sxl)	2
Cactus	cactus (cact)	2
Torso	torso (tor)	2
EGF-Receptor	EGF receptor (Egfr)	2
Daughterless	daughterless (da)	2
extra sex combs	extra sexcombs (esc)	2
TmI	Tropomyosin 1 (Tm1)	2
NINAC	neither inactivation nor afterpotential C (ninaC)	2
Grk	gurken (grk)	1
lethal-of- scute	lethal of scute (l(1)sc)	1
alpha- spectrin	alpha Spectrin (alpha-Spec)	1
TOLLOID	tolloid (tld)	1
Ultra-bithorax	Ultrabithorax (Ubx)	1
tropomyosin I	Tropomyosin 1 (Tm1)	1
vgBG	vestigial (vg)	1
Lethal of Scute	lethal of scute (l(1)sc)	1
Scute	scute (sc)	1
Sevenless	sevenless (sev)	1
suppressor of Hairy-wing	suppressor of Hairy wing (su(Hw))	1
ANTP	Antennapedia (Antp)	1
TmII	Tropomyosin 2 (Tm2)	1
Protein Kinase A	Protein Kinase A (PKA)	1
serendipity delta	Serendipity delta (Sry-delta)	1
acetylcholinesterase	Acetylcholine esterase (Ace)	1
Segregation Distorter	Segregation distorter (Sd)	1
SCW	screw (scw)	1
troponin I	wings up A (wupA)	1
Hunchback	hunchback (hb)	1
Runt	runt (run)	1
Abdominal- A	abdominal A (abd-A)	1
Zeste-White 3	shaggy (sgg)	1
histone H1	Histone H1 (His1)	1
AceIJ29	Acetylcholine esterase (Ace)	1
AceIJ40	Acetylcholine esterase (Ace)	1
Hsp90	Heat shock protein 83 (Hsp83)	1
EMS	empty spiracles (ems)	1
double sex	doublesex (dsx)	1
absent, small or homeotic discs1	absent, small, or homeotic discs 1 (ash1)	1
histone H3	Histone H3 (His3)	1
D- Mek	Downstream of raf1 (Dsor1)	1
PSI	P-element somatic inhibitor (Psi)	1

Label	Gène	Fréquence
bigbrain	big brain (bib)	1
TRA	transformer (tra)	1
extra-macrochaete	extra macrochaetae (emc)	1
Achaete	achaete (ac)	1
Beta Tub56D	betaTubulin56D (betaTub56D)	1
Beta 3 tubulin	betaTubulin60D (betaTub60D)	1
Phosrestin II	Arrestin A (Arr1)	1
Zeste-white 3	shaggy (sgg)	1

Tableau 37 Définitions aberrantes

Certaines définitions sont manifestement inappropriées pour identifier des gènes à l'intérieur de texte. Le label de la première colonne est sensé désigner le gène de la deuxième colonne. Plus de 4500 définitions de cette sorte ont été dénombrées. Nous listons ici celles qui concernent les gènes les plus répandus.

Label	Gène
l(2)22Fa	decapentaplegic (dpp)
Hin-d	decapentaplegic (dpp)
ho	decapentaplegic (dpp)
Haplo-insufficient	decapentaplegic (dpp)
l(2)10638	decapentaplegic (dpp)
M(2)23AB	decapentaplegic (dpp)
M(2)LS1	decapentaplegic (dpp)
I	wingless (wg)
Complementation group I	wingless (wg)
l(2)02657	wingless (wg)
l(2)rO727	wingless (wg)
bx	Ultrabithorax (Ubx)
bithorax	Ultrabithorax (Ubx)
l(3)89Eb	Ultrabithorax (Ubx)
prd4	bicoid (bcd)
PRD gene 4	bicoid (bcd)
mat(2)dorsal	dorsal (dl)
l(1)6Fa	Sex lethal (Sxl)
Fl	Sex lethal (Sxl)
Female lethal	Sex lethal (Sxl)
l(1)3Cb	Notch (N)
l(1)N	Notch (N)
T5	achaete (ac)
l(3)08247	hairy (h)
l(3)85Ah	hunchback (hb)
l(3)rM384	hairy (h)
transcript group V	engrailed (en)
V	engrailed (en)
Humeral	Antennapedia (Antp)
l(3)84Ba	Antennapedia (Antp)
ANTC	Antennapedia (Antp)
Hu	Antennapedia (Antp)
l(2)57Ea	EGF receptor (Egfr)
l(2)57DEFa	EGF receptor (Egfr)
l(2)57EFa	EGF receptor (Egfr)
l(1)IV	extradenticle (exd)
Drosophila epidermal growth factor receptor homologue	EGF receptor (Egfr)
T4	scute (sc)
l(1)1Ba	scute (sc)
l(2)br28	snail (sna)
l(3)br28	snail (sna)
l(2)35Db	snail (sna)
l35Db	snail (sna)
br28	snail (sna)
l(3)84Ag	fushi tarazu (ftz)
l(3)07117	nanos (nos)
l(3)j6E3	squid (sqd)
l(2)49Ea	Posterior sex combs (Psc)
l(2)vr14	Posterior sex combs (Psc)
vr14	Posterior sex combs (Psc)

Tableau 38 Liste des tables présentes dans la base de données
La deuxième colonne renvoie au code de la section qui décrit la table.

Intitulé	Section	Contenu de la table
AR	Chapitre 2I.A.2.a.ii.1er)	Annotateur Résumé. Personne qui a annoté le résumé
CD	Chapitre 2I.A.3.c.iii	Confiance dans la Définition. Type de traitement à donner à la définition en fonction de la confiance qu'on lui porte.
CI	Chapitre 2II.A.1	Couple de gènes en Interaction. Reconnaissance d'interaction.
DG	Chapitre 2I.A.3.c.i	Définition de gène. Association entre un gène et un label.
EB	Chapitre 2I.A.3.a.ii.1er)	Entité Biologique. Rubrique du dictionnaire des gènes.
FGA	Chapitre 2I.A.3.a.iii	Filiation Gène ou Assimilé. Table de relation père-fils pour les gènes.
GA	Chapitre 2I.A.3.a.i	Gène ou objet assimilé.
IPF_PF_IRM_	Chapitre 2III.C	Phrase issue de Flybase.
Lm	Chapitre 2II.B.1.a	Lemme. Forme lemmatisé des termes spécifiques de l'interaction.
MB	Chapitre 2II.B.1.b	Mot brut. Forme non lemmatisé des mots spécifiques de l'interaction
NG	Chapitre 2I.A.3.b.i	Nom de Gène. Label, terme utilisé pour désigner un gène.
OED	Chapitre 2I.A.3.c.ii.2e)	Origine de l'Enregistrement du Dictionnaire. Origine de la définition de gène.
OI	Chapitre 2II.A.2	Ordre dans l'Interaction.
ORM	Chapitre 2I.A.2.a.ii.2e)	Origine du Résumé Medline.
PGA	Chapitre 2I.A.3.a.ii.2e)	Provenance Gène ou Assimilé.
PM	Chapitre 2I.A.2.b	Phrase Medline.
PRG	Chapitre 2I.A.4.b.ii	Processus de Reconnaissance des Gènes. Etape dans le processus de reconnaissance des gènes dans les textes.
PRI	Chapitre 2II.A.3	Processus de Reconnaissance des Interactions. Type de reconnaissance pour les interactions.
RDG	Chapitre 2I.A.4.b.i	Reconnaissance de Définition de Gène. Identification d'une définition de gène en un lieu donné d'un texte.
RM	Chapitre 2I.A.2.a.i	Résumé Medline. Le résumé en entier.
RMB	Chapitre 2II.B.2	Reconnaissance Mot Brut. Identification d'un mot brut dans une phrase en un lieu donné.
RNG	Chapitre 2I.A.4.a	Reconnaissance Nom de Gène. Identification d'un label dans une phrase à une certaine position.
RNGR	Chapitre 2I.A.3.b.ii	Reconnaissance Nom de Gène Réflexive. Identification d'un label dans un autre label en une position donnée.
RT	Chapitre 2I.A.3.b.vi	Relation de Transformation. Type de transformation qui mène d'un label à un autre label.
TNG	Chapitre 2I.A.3.b.vi	Transformation Nom de Gène. Relation de transformation entre les labels.
TR	Chapitre 2I.A.3.b.v	Type de Reconnaissance. Catégorie de labels et donc traitement approprié à faire lors de la reconnaissance.
Ty	Chapitre 2I.A.3.c.ii.1er)	Type de la définition. Catégorie Nom abrégé, Nom Complet, Synonyme...

Chapitre 2 Mise en œuvre

Nous traitons dans cette partie de la mise en œuvre de l'identification des gènes et de la reconnaissance des interactions.

I. MISE EN ŒUVRE DU PROGRAMME D'IDENTIFICATION DES GÈNES

Nous traitons dans cette partie de l'outil et des méthodes que nous avons mis en œuvre pour résoudre le problème de l'identification des gènes. Dans un premier temps, nous exposerons la façon dont les informations nécessaires à l'analyse sont représentés dans la base de données. Dans un deuxième temps nous expliquerons comment le système utilise les données de façon à parvenir à identifier les gènes par l'analyse des textes. Dans un troisième temps, nous expliquerons comment nous avons collecté et structuré les données de façon à les faire rentrer dans la base de données tel quelle est structurée.

A. STRUCTURE DE DONNÉES POUR L'IDENTIFICATION DES GÈNES DANS LES TEXTES

Dans cette partie, nous traitons de la façon dont les informations nécessaires à l'identification des gènes sont structurées à l'intérieur de la base de données. La méthode que nous avons suivi pour mettre les données sous cette forme ne pourra logiquement être donnée que plus tard. Cela sera traité en C. De même, la façon dont l'outil fonctionne pour identifier les gènes ne pourra être expliquée qu'après cette partie. Cela sera fait en B.

1. Préliminaires

Avant d'aborder la structure de la base de données, nous donnons dans cette partie quelques éléments sur les bases de données relationnelles qui seront utiles tout au long de l'exposé.

a. Notions sur les bases de données relationnelles

Les données contenues dans la base de données sont organisées en tables. Nous pouvons représenter ces tables par des tableaux. Dans ce cas, chaque ligne représente un enregistrement (ou individu) et chaque colonne représente un champ (ou caractéristique). A l'intersection d'une ligne et d'une colonne se trouve la valeur du champ colonne pour l'enregistrement ligne.

Les enregistrements d'une table sont généralement munis d'un numéro unique. Le champ correspondant est nommé ***clef primaire***.

Dans un champ d'une table, il arrive fréquemment que l'on fasse référence à un enregistrement d'une autre table. Dans ce cas on utilise précisément la ***clef primaire*** de la deuxième table pour indiquer de façon unique l'enregistrement en question. On dira dans ce cas que le champ de la première table est une ***clef externe***. On dira qu'il y a une relation entre les deux tables. La plupart du temps cette relation est une relation « de un à plusieurs ». Cela signifie que plusieurs enregistrements de la première table peuvent faire référence au même enregistrement de la deuxième table, mais que chaque référence à un enregistrement est univoque. L'ensemble des liens présents dans la base de données est présenté Figure 5.

b. Conventions sur les noms de champs et de tables

Dans la base de données nous utilisons des abréviations pour désigner les tables. Ces abréviations sont constituées en général de deux à quatre lettres en majuscule. Cela s'avère pratique pour que les requêtes ne soient pas constituées d'un texte trop long de façon à rester lisible. La liste des tables est présentée dans le tableau 38.

Pour les ***clefs primaires***, nous utiliserons toujours une abréviation qui commence par la lettre i comme identifiant.

c. Quelques principes sur la structuration des données

i. Les garanties d'intégrité des données

Le gestionnaire de base de données peut prendre en charge l'intégrité des donnés, c'est à dire leurs cohérences. Il suffit pour cela d'édicter des règles d'intégrité des données. Le gestionnaire de base de données va alors empêcher toute opération qui aboutirait à une violation de ces règles. Voyons maintenant quelle sont les principales règles.

1er) Garantir la présence des enregistrements cités dans une table

Quand une table fait référence à des enregistrements d'une autre table, il est important que les enregistrements cités existent bel et bien. Cette règle s'appelle ***l'intégrité relationnelle***. Cette règle est mise en danger quand on supprime un enregistrement. Pour la garantir, le système peut prendre l'initiative de supprimer en cascade les enregistrements qui font référence à l'enregistrement en question. C'est une règle que nous avons adoptée systématiquement sauf dans le cas où cela pourrait être dangereux pour un utilisateur non averti. En particulier nous ne l'avons pas fait pour les liens vers des « petites » tables qui ne font en fait que donner les quelques modalités, une douzaine au plus, qui sont possibles pour la valeur d'un champ.

2e) Garantir l'absence de doublons dans les enregistrements

Les doublons sont en général proscrits dans les bases de données relationnelles. Un ***doublon*** est un enregistrement en double dans une table. Dans certains cas, on ne considère que certains champs pour juger de la présence de doublons. Pour garantir l'absence de doublons dans une table, des index sont posés sur un ou plusieurs champs de la table. Un index est une structure interne au système de gestion de base de données qui donne pour chaque valeur du champ ou des couples de champs qui définissent l'index, un accès direct aux enregistrements qui utilisent cette valeur. Pour garantir l'absence de doublons, on pose comme condition que l'index doit être sans doublons. L'absence de doublons dans un champ est mise en péril par l'ajout d'enregistrements. Le système, pour maintenir l'intégrité des données, va donc ignorer les tentatives d'ajouts d'enregistrements qui aboutiraient à la création de doublons.

ii. Structure des données pour permettre les mises à jour.

Les données contenues dans la base de données sont pour l'essentiel issues d'autres bases de données, à savoir *Medline* et *Flybase*. Elles ont été importées par des processus automatiques. La base est conçue pour pouvoir permettre des "imports" successifs. Il est important de faire en sorte que les données dont nous avons décidé, pour une raison ou

pour une autre, de ne pas tenir compte, ne soient pas purement et simplement supprimées. En effet, si tel était le cas, ces données risqueraient d'être réintroduites par erreur dans la base lors de la prochaine mise à jour de la base de données. Il est impératif de garder une trace des données que nous avons souhaité de mettre "hors jeu", la suppression n'étant pas une solution acceptable.

Pour permettre la mise à jour, nous avons introduit à chaque fois que nécessaire, des champs de d'activation ou d'inactivation. Par défaut, tous les enregistrements sont actifs. Les enregistrements indésirables sont inactivés grâce à ce champ.

2. Structure de données pour les textes

a. Structure de données pour les résumés

i. La table des résumés

Les 744 résumés que compte la base de données sont contenus dans une table. Le tableau 39 présente un exemple d'enregistrement de cette table. La table est intitulée *RM* pour *Résumé Medline.*.

Tableau 39 Table des résumés
Est présenté ici un enregistrement parmi d'autre de la table des résumés Medline.

Champ	Signification	Contenu
IRM	Clef primaire	94326643
RM	Texte du résumé	Successive alternative cell fate choices in the imaginal disc epithelium lead to the differentiation of a relatively invariant pattern of multicellular adult sensory organs in Drosophila. We show here that ...
IO	Origine	Lié à Flybase
PIR	*IVI*	-0,07
PIR1	*IVI* variante	-0,04
Vu	Annoté entiérement	Oui
AIR	Alias Numéro	455
IA	Annotateur	Bernard 2
Date	Date de création de l'enregistrement	21/04/00 16:27:55

Le numéro utilisé comme clef primaire n'est autre que le numéro utilisé par la base de données *Medline*. Le champ *AIR* fournit un deuxième système de numérotation, qui est propre à la base de données. Il s'agit du numéro d'ordre dans la présentation des résumés dans le formulaire d'annotation. Le texte du résumé a été tronqué pour améliorer la présentation. Deux champs sont consacrés à l'*IVI* de façon à permettre une comparaison entre différentes méthodes de calcul de cet indice. Nous n'avons pas utilisé l'intitulé *IVI* car le préfixe I est réservé aux clefs. Le champ *Vu* permet de savoir si le résumé à été annoté manuellement à la fois pour la reconnaissance des définitions de gènes et pour l'identification des interactions.

ii. Les tables annexes

1er) Structure de données pour le suivi de l'annotation

L'annotation s'est faite en plusieurs temps et par différentes personnes. Pour mémoriser ce qui a été fait sur chaque résumé, un champ dans le résumé est consacré à cette information.

Ce champ intitulé *IA* contient une référence à un enregistrement de la table des annotateurs (AR). Le contenu de cette table est donné dans le tableau 40.

Tableau 40 Table des annotateurs
Cette table permet le suivie du processus d'annotation. La personne ainsi que le degré d'achèvement de la notation y est inscrit.

Clef	Annotateur	Description
1	à faire...	
2	Ambroise 2	Interaction, avec le sens. Reconnaissance des gènes, avec ajout des gènes manquants, sans les groupes
3	Bernard 2	Interaction, avec le sens. Reconnaissance des gènes, avec ajout des gènes manquants, sans les groupes
4	Violaine 3	
5	Denys 1	Interaction, avec le sens. Reconnaissance des gènes, avec ajout des gènes manquants, sans les groupes
6	Violaine et Deny 1	Interaction, avec le sens. Reconnaissance des gènes, sans ajout des labels manquant, sans les groupes
7	Ambroise 1	Interaction, avec le sens. Reconnaissance des gènes, sans ajout des labels manquant, sans les groupes
8	Bernard 1	Interaction, avec le sens. Reconnaissance des gènes, sans ajout des labels manquant, sans les groupes
9	Ambroise 3	Interaction, avec le sens. Reconnaissance des gènes, avec ajout des gènes manquants, avec les groupes
10	Bernard 3	Interaction, avec le sens. Reconnaissance des gènes, avec ajout des gènes manquants, avec les groupes
11	Ambroise 0	Interaction sans le sens. Pas de validation de la reconnaissance des gènes par les labels

2e) Structure de données pour l'origine de l'enregistrement

La base de données contient actuellement deux types distincts d'enregistrements. Le tableau 41 donne la description de ces deux types. La table est intitulée *ORM* pour *Origine Résumé Medline.*.

Tableau 41 La table des origines des résumés
Les résumés de la base sont de deux types. Chaque enregistrement de la table représente un type.

IO	Origine	Commentaires
1	Lié à Flybase	Les résumés de ce type sont cités par au moins une phrase de l'échantillon issue de Flybase
2	Sans nom de gène	Les résumés de ce type n'ont pas en principe de nom de gène. Ils ont été choisits spécialement pour cette propriété.

b. Structure de données pour les phrases qui constituent les résumés

Les phrases qui composent les résumés sont représentées dans une table distincte de celle des résumés. Le tableau 42 donne un exemple d'enregistrement de cette table. La table est intitulée *PM* pour *Phrase Medline.*.

Tableau 42 Table des phrases extraites de Medline
Le tableau donne exemple d'enregistrement.

Champ	Contenu
Clef Phrase	21794
Phrase	These results, along with the intermediate SOP phenotype observed in Suppressor of Hairless; Hairless double mutant imaginal discs, suggest that the two genes act antagonistically to commit imaginal disc cells stably to alternative fates
Remarque Vu	Non
N° Résumé	94326643
N° d'ordre	8
IVI	0,06
IVI bis	0,09
Problème	Partenaires non identifiés
Mauvaise sègmentation	Nom
Gène Absent	Nom
Date	21/04/00 16:57:57

Les phrases d'un même résumé sont numérotées grâce au champ *Numéro d'ordre*.

Deux champs sont destinés à recevoir chacun une valeur de l'*IVI* de façon à permettre les comparaisons entre plusieurs méthodes.

Quand une erreur s'est produite dans l'opération de segmentation du résumé en phrases, le champ *mauvaise segmentation* prend la valeur *Oui*.

Le champ *Gène Absent* a été utilisé à un moment où la structure de la base de données n'était pas achevée. Les informations qu'il contient actuellement sont destinées à être transférées dans la table *CI* des couples de gènes en interaction.

Le champ *Problème* reçoit les remarques faites sur la phrase lors de l'annotation. Lors du passage en revue de tous les problèmes rencontrés, le champ *Remarques Vu* est coché.

3. Structure de données pour le dictionnaire des gènes

a. Structure de données pour les gènes ou objets assimilés

i. La table des gènes ou objets assimilés

Les gènes ainsi que les objets qui peuvent leurs être assimilés sont contenus dans la table ***gènes et assimilés*** (GA) dont un enregistrement est présenté dans le tableau 43.

Tableau 43 Table des gènes ou objets assimilés
Le tableau donne un exemple d'enregistrement de la table.

Champ	Contenu
Clef Gène	13
Objet	abdominal A (abd-A)
Symbole	abd-A
N° Flybase	14
Classe	Gène
Validation	Actif
Problème	
Provenance	Flybase
Référant	abdominal A (abd-A)
Date	21/04/00 17:10:05

Le numéro du gène dans la base de données *Flybase* est conservé mais ne constitue pas la clef primaire. En effet, lors de l'introduction de nouveaux enregistrements, on ne peut pas garantir facilement que l'on n'utilise pas des numéros identiques à ceux que *Flybase* utilise pour ajouter de son coté, de nouveaux enregistrements.

Le champ légendé *objet* permet de donner un nom convivial à l'enregistrement. Au départ il est créé à partir du *nom complet* et du *symbole* du gène, mais il est librement modifiable par la suite.

Le champ *symbole* est une redite par rapport à l'information qui se trouve dans la table des définitions de gènes, mais sa présence est néanmoins utile car elle évite de faire référence sans cesse à cette table, ce qui aurait pour conséquence de ralentir l'affichage des formulaires de consultation de la base de données. Il est rempli au départ à partir des informations contenues dans la table des définitions.

La *classe* permet de gérer les rubriques du dictionnaire. Cette structure permet de traiter de la même façon des objets qui ne sont pas des gènes mais qui jouent le même rôle dans notre base de données.

Le champ *référant* permet de faire référence au gène père dans la relation d'allélisme. Un gène qui n'est l'allèle d'aucun autre gène fera référence à lui-même, comme c'est le cas dans l'exemple présenté dans le tableau 43.

Le champ *validation* permet d'invalider des enregistrements sans les supprimer de façon à permettre la mise à jour des données avec la base de données *Flybase* qui est la source des données. Seuls quelques enregistrements ont été invalidés car ils formaient des doublons.

ii. Les tables annexes à la table des gènes

1er) Structure de données pour les rubriques du dictionnaire

Le dictionnaire est organisé en plusieurs rubriques selon la classe de l'objet considéré. La table ***Entité Biologique*** (EB) fait l'inventaire de toutes les modalités du champ *classe*. Le tableau 44 donne le contenu de cette table dans son exhaustivité.

Tableau 44 Table des entités biologiques
Les modalités du champ classe sont données dans cette table.

Clef	Classe
1	Gène
2	Famille de protéines
3	Complexe de protéines
4	Complexe de gènes
5	Termes spécifiques
6	Famille
7	Allèle
9	Gène mammifères

2e) Structure de données pour la gestion de la provenance du gène

Les gènes ou objets assimilés sont pour la plupart issus de *Flybase* mais un petit nombre d'entre eux ont été introduits manuellement lors de l'annotation. La table *Provenance des Gènes ou objets Assimilés* (*PGA*) dont le contenu est donné dans le tableau 45 est prévue à cet effet.

Tableau 45 Table Provenances des gènes
La table est donnée ici dans son intégralité.

Clef	Origine Gène
1	Flybase
2	Ajout

iii. Structure de données pour la gestion de la filiation

Les nouvelles entités biologiques que sont les complexes de gènes, les complexes de protéines ou les familles de protéines sont liées aux gènes déjà présents par une relation d'appartenance. L'ensemble de ces liens est contenu dans la table dite de *filiation*. Le tableau 46 donne un extrait de cette table. Cette table est intitulée *FGA* pour *Filiation Gène ou Assimilé*.

Tableau 46 Table des filiations
La table complète compte 57 lignes. Nous en présentons ici un extrait.

Clef Filiation	Objet fils	Objet père	Classe du père
4	abdominal A (abd-A)	bithorax complex (BXC)	Complexe de gènes
5	Abdominal B (Abd-B)	bithorax complex (BXC)	Complexe de gènes
6	Ultrabithorax (Ubx)	bithorax complex (BXC)	Complexe de gènes
95	sloppy paired 2 (slp2)	Sloppy paired (slp)	Complexe de gènes
94	sloppy paired 1 (slp1)	Sloppy paired (slp)	Complexe de gènes
133	Antennapedia (Antp)	Antennapedia complex (ANT-C)	Complexe de gènes
143	Polycomb (Pc)	Polycomb group (Pc-G)	Complexe de protéines
2	Actin 42A (Act42A)	Actin	Famille de protéines
20	Actin 88F (Act88F)	Actin	Famille de protéines
19	Actin 87E (Act87E)	Actin	Famille de protéines
18	Actin 79B (Act79B)	Actin	Famille de protéines
17	Actin 57B (Act57B)	Actin	Famille de protéines
1	Actin 5C (Act5C)	Actin	Famille de protéines

b. Structure de données pour les labels

i. La table des labels

Les labels, c'est à dire les chaînes de caractères qui dans un texte peuvent désigner un gène ou un objet assimilé, sont répertoriés dans une table. Le tableau 47 donne un exemple d'enregistrement de la table. La table est intitulée *NG* pour *Nom de Gène*..

Tableau 47 Table des labels
Un exemple d'enregistrement est donné dans ce tableau.

Champ	Contenu
Clef Label	108
Label	AbdA
Transcode	. Abd . A
Reconnaître	Non renseigné
Nouveau	Non
Date	21/04/00 16:20:45

L'utilité des champs *Transcode*, *Reconnaître* et *Nouveau* seront expliqués dans les sections qui suivent.

ii. Structure de données pour la relation d'inclusion

Les labels sont inclus les uns dans les autres et cette inclusion est une information essentielle au processus d'identification des gènes dans les textes. Nous avons discuté de ce point dans la section Chapitre 1I.B.5. Ces informations sont consignées dans la table d'*inclusion*. Le tableau 48 donne un exemple d'enregistrement de cette table. La table est intitulée *RNGR* pour *Reconnaissance de Nom de Gène Réflexive*..

Tableau 48 Table des inclusions
Un exemple d'enregistrement de la table est donné dans ce tableau. Le label *Hairless* est inclue en position 15 dans le label *Suppressor of Hairless*.

Champ	Contenu
Label	Suppressor of Hairless
Inclue	Hairless
Position	15
Date	07/06/00 21:25:48

La *position* est le numéro du caractère où commence le label inclus dans la chaîne de caractères qui constitue le texte de la phrase.

iii. Structure de données pour faciliter l'actualisation des données

La table des inclusions est très volumineuse et la construire demande beaucoup de temps à la machine (quatre nuits). Elle ne peut donc pas être reconstruite à chaque fois que l'on souhaite lancer le processus d'identification des gènes dans des textes. Pour éviter cette reconstruction totale, une reconstruction partielle ou plutôt une complémentation a été imaginée. Il s'agit de ne reconstruire la table d'inclusion que pour les enregistrements nouvellement introduits dans la table des labels. Le champ nouveau a été introduit à cet effet dans la table des labels. Il prend par défaut la valeur *Oui* lors de la création d'un nouvel enregistrement. Lors de la complémentation de la table d'inclusion, seuls les enregistrements dont le champ nouveau prend la valeur *Oui* sont pris en compte. A la fin de la complémentation le champ nouveau est mis à *Non* pour tous les enregistrements de la table des labels. La complémentation elle-même de la table d'inclusion sera expliquée dans la section C.2.d.

iv. La garantie de l'unicité

Il est important qu'il n'y ait pas de doublons dans les tables. Les doublons nuissent à la bonne interprétation des données et notamment lors de l'établissement de statistiques. Pour la table des labels, il est plus difficile d'assurer cette exigence. En effet, il est possible de poser des règles d'intégrité des données dont le logiciel de gestion de base de données va ensuite assurer de lui-même. Ainsi, après avoir fait en sorte qu'il n'y a pas initialement de doublons dans une table, on peut interdire leur existence pour la suite. Dès lors, le SGBD va faire en sorte de ne jamais en créer en refusant tout nouvel enregistrement qui serait en contradiction avec cette règle. Pour imposer l'absence de doublons dans un champ, on demande à ce qu'il soit indexé sans doublons.

Pour la table des labels, il est plus difficile de garantir l'absence de doublons. Il n'est pas possible d'imposer que le champ *label* de la table ne contienne pas de doublons car le logiciel que nous utilisons ne fait pas la différence entre les majuscules et les minuscules lors de la construction des index. Le champ *label* est donc doublé d'un autre champ appelé *transcode*. Ce champ contient la même information mais avec un codage particulier qui permet au logiciel de faire la différence entre les majuscules et les minuscules. Le transcodage consiste à insérer un caractère (par ailleurs absent des données) devant chaque majuscule. L'exigence d'unicité de l'enregistrement est posée dans le champ *transcode*.

v. Structure de données pour la caractérisation du type de traitement à faire sur chaque label

Chaque label peut être traité de façon différente par le système d'identification des gènes en fonction de caractéristiques propres au label. La table qui contient cette information s'intitule *TR* pour *Type de reconnaissance*. Les informations qu'elle contient ont été mises dans deux tableaux, car elles ne tenaient pas dans un seul. Le tableau 49 donne la liste des catégories et pour chaque catégorie, le traitement qui est réservé aux labels appartenant à cette catégorie. Le tableau 50 commente chaque catégorie en précisant le type de labels qu'elle contient.

Tableau 49 Table type de reconnaissance (première partie)
Les labels sont classés en fonction du type de traitement à effectuer lors du processus d'identification des gènes dans les textes. Les catégories de labels sont données dans ce tableau.

Clef	Reconnaissance	Indexer	2ième Vague	Désindexer si début	Interpréter	Confirmer	Exemples
6	Abérant	Non	Non	Non	Non	Non	Tableau 37

Clef	Reconnaissance	Indexer	2ième Vague	Désindexer si début	Interpréter	Confirmer	Exemples
2	Mot vide si début de phrase	Non	Oui	Oui	Oui	Oui	

Tableau 76 Les contradictions du dictionnaire. Les contradictions présentent dans le dictionnaire issu de *Flybase* sont présentés dans ce tableau. Le label intervient dans les définitions de chacun des deux gènes.

Label	Gène 1	Gène 2
ACE1	Amplification-control-element-on-1 (ACE1)	Chorion protein
ACE3	Amplification-control-element-on-3 (ACE3)	Chorion protein
Als	alae sublatae (als)	nicotinic Acetyl (nAcRa
And	Androcam (And)	dusky (dy)
Ang	angle wing (ang)	anomogenitals (
angle wing	angle wing (ang)	angle winglike (
Ant	antennaless (ant)	empty spiracles
Apa	Apigmented abdomen (Apa)	Saposin-related
ARS	Autonomously Replicating Sequence (ARS)	Arylsulfatase (A
Aurora	aurora (aur)	aurora transpos
Bald	bald (bld)	balding (bd)
Bam	bag of marbles (bam)	breaks at metap
Bb	Bubble (Bb)	lethal (2) 37Bb
Bb	bobbed (bb)	Y-bobbed (Ybb

Clef	Reconnaissance	Indexer	2ième Vague	Désindexer si début	Interpréter	Confirmer	Exemples
1	Mot vide	Non	Oui	Non	Oui	Oui	Tableau 78
11	Ambigu en début de phrase	Oui	Non	Oui	Oui	Non	Tableau 79
10	Trop ambigu	Oui	Non	Non	Non	Non	Tableau 84
13	Terme spécifique	Oui	Non	Non	Non	Non	Tableau 80
3	Plutôt ambigu	Oui	Non	Non	Oui	Oui	Tableau 81
4	Peut-être ambigu	Oui	Non	Non	Oui	Oui	Tableau 82
8	Désambiguïsation en cours	Oui	Non	Non	Oui	Oui	Tableau 83
12	Ambiguité constatée mais marginale	Oui	Non	Non	Oui	Oui	Tableau 84
0	Non renseigné	Oui	Non	Non	Oui	Non	
5	Peu ambigu	Oui	Non	Non	Oui	Non	Tableau 85
7	Désambigüé	Oui	Non	Non	Oui	Non	Tableau 84
9	Spécifié univoque	Oui	Non	Non	Oui	Non	Tableau 86

Tableau 50 Table Type de reconnaissance (deuxième partie)
Le champ commentaire de la table permet de préciser l'intitulé de l'enregistrement.

Clef	Reconnaissance	Commentaire
6	Abérant	Label abérant. Ne sera pas reconnu. Ex : gene 1
2	Mot vide si début de phrase	Mot très courant si en première position dans la phrase, sinon plutôt rare. Ex : We
1	Mot vide	Mot très courant de l'anglais. Exemple : at. On les indexes que dans les résumés où les gènes associés semblent présents.
11	Ambigu en début de phrase	Mot de début de phrase assez courant en anglais. Ex : Midway
10	Trop ambigu	Mot très courant de l'anglais et que l'on ne prendra pas en compte
13	Terme spécifique	Elément du lexique qui n'est pas le nom d'un gènes mais qui inclus un label
3	Plutôt ambigu	Mot trop courant de l'anglais et que l'on ne prendra en compte que s'il y a confirmation
4	Peut-être ambigu	Mot possiblement ambigu. A voir plus tard. Interprété pour l'instant.
8	Désambiguïsation en cours	Désambiguïsation en cours. Pour l'instant le label n'est pas interprété.
12	Ambiguité constatée mais marginale	On a trouvé un ou plusieurs cas où le label prétait à confusion mais cela semble marginale
0	Non renseigné	La nature du label n'a pas été étudié.
5	Peu ambigu	Mot moins ambigu qu'il n'y parait. Il faut le reconaître.
7	Désambigüé	Label au dépard ambigu pour lequel on a une solution de désambiguïsation. Ex : dorsal
9	Spécifié univoque	Label considéré comme univoque par l'annotateur. Recommandé pour les labels rentré manuellement.

Les étapes du traitement seront expliquées dans la partie sur la méthode de reconnaissance des gènes dans les textes. A ce stade, la seule chose à savoir est que les types de traitement à faire sur les labels sont indiqués dans la table *type de reconnaissance*.

La liste des termes appartenant à chaque catégorie est donnée dans la plupart des cas en annexe. Le numéro du tableau à voir est indiqué dans la dernière colonne. Dans certains cas, il s'agit d'un tableau que nous avons donné dans la partie sur l'analyse du problème de l'identification des gènes.

La modalité *non renseigné* est utilisée au cours du processus d'annotation, mais à l'issue de celui-ci aucun label n'a ce type de reconnaissance.

vi. Structure de données pour la gestion de la relation de transformation

Les labels sont liés entre eux par des relations de transformation comme nous l'avons vu dans la section Chapitre 1I.D.2. Des exemples sont donnés du tableau 20 au tableau 23. La table des transformations *TNG,* en abrégé pour *transformation nom de gène*, fait l'inventaire de toutes ces relations entre labels. Le tableau 51 donne un exemple d'enregistrement de cette table.

Tableau 51 Table des transformations
Ce tableau donne un enregistrement parmi d'autre de la table. Les labels 1 et 2 sont liés par une relation de transformation.

Champ	Contenu
Label 1	wingless
transformation	1ière lettre en majuscule
Label 2	Wingless

Les modalités possibles pour la relation de transformation sont listées dans la table *Relation de transformation* (*RT*) qui est donnée dans le tableau 52.

Tableau 52 Table des relations de transformations
Cette table donne les cinq types possibles de relation de transformation.

Clef	Relation	Commentaire
1	Tout en minuscules	Les lettres sont misses en minuscules
2	Tout en majuscules	Les lettres sont misses en majuscules
3	1ière lettre en majuscule	La première lettre est mise en majuscule
4	espace -> tiret	chaque espace est remplacé par un tiret
5	tiret -> espace	chaque tiret est remplacé par un espace

c. Structure de données pour les définitions

i. La table des définitions de gènes

Le dictionnaire des gènes et objets assimilés est structuré à l'aide de définitions comme expliqué en Partie 2Chapitre 1I.B.2. Un exemple d'enregistrement de la table des *définitions de gène* (DG) est donné dans le tableau 53. La signification des champs *Type*, *Origine* et *Confiance* sont expliqués dans les sections qui suivent.

Tableau 53 Table des définitions
Un exemple d'enregistrement de la table est donné dans ce tableau.

Champ	Contenu
Clef Définition	10856
Label	abdA
Type	Synonyme
Origine	Flybase
Confiance	correcte
Problème	
Date	21/04/00 16:16:49

ii. Les tables annexes

1er) La table des types de définition

Les différents types de définitions sont énumérés dans la table *type de définition*. Le contenu de cette table est donné dans le tableau 54. La table est intitulée *Ty*.

Tableau 54 Table des types de définition
Cette table fait l'inventaire des modalités possible pour le champ type de la table définition.

Clef	Type
1	Symbole
2	Nom Complet
3	Synonyme
4	Protéine

La valeur *protéine* est utilisée pour les définitions qui associent un nom de protéine au gène qui code pour cette protéine. Dans *Flybase* cette modalité est absente. Dans notre base de données, les définitions de type protéine sont en général issues du processus d'ajout automatique de définition qui sera expliqué en Partie 2Chapitre 2I.C.3.a.

2e) Structure de données pour le suivi de l'origine des définitions

Les modalités possibles du champ *Origine* de la table des définitions sont données dans la table *Origine des définitions* (appelé *OED* pour *Origine des Enregistrements du Dictionnaire*) dont le contenu est donné dans le tableau 55.

Tableau 55 Table *origine des définitions*
Cette table fait l'inventaire des modalités possible du champ origine de la table définition.

Clef	Origine	Confiance	Variante
1	Flybase	Oui	Non
2	Ajout Manuel	Oui	Non
3	Tout en majuscule	Non	Oui
4	1ère lettre en majuscule	Non	Oui
5	Tout en minuscule	Non	Oui
6	Tiret -> espace	Non	Oui
7	Espace -> tiret	Non	Oui
8	réimport	Non	Non

Il y a deux types principaux d'origines :

- Les origines de type *variantes* (de 3 à 7 dans la table) correspondent à des définitions qui ont été ajoutées automatiquement. C'est ce que nous avons appelé des ***définitions variantes***.
- Les origines de type *confiance* qui correspondent à des définitions qui, soit étaient présentes dans *Flybase*, soit ont été rajoutées manuellement par l'annotateur.

L'origine *réimport* correspond à des enregistrements qui avaient été supprimés à un stade précoce du traitement, car ils ne sont pas appropriés pour notre travail. Cela est expliqué dans la section Partie 2Chapitre 1I.E.2. Ces enregistrements de la table *définition* ont été réimportés dans la base pour permettre une mise à jour des données à l'avenir. Ces enregistrements sont disqualifiés à toutes les étapes de l'identification des gènes par la présence de la modalité *désactivée sur liste* dans le champ *confiance*.

iii. Structure de données pour gérer la confiance mise dans les définitions

Les modalités possibles du champ *confiance* de la table *définitions* sont données dans la table *Confiance dans les Définitions* (CD) qui est donnée intégralement dans le tableau 56.

Tableau 56 Table *confiance dans les définitions*
Cette table fait l'inventaire des modalités possible du champ confiance de la table des définitions.

Clef	Confiance	Valide	Prendre	Confirmation	Commentaire
1	correcte	Oui	Oui	Non	La définition n'est pas contredite par une définition valide.
2	privilégiée	Oui	Oui	Non	La définition est contredite mais elle reste crédible.
3	à confirmer	Oui	Oui	Oui	La définition est contredite par une définition plus crédible. Elle n'est pas prise en compte à moins qu'elle soit confirmée par ailleurs.
5	imprécise	Oui	Non	Non	Le label ne désigne pas un gène précis mais une collection de gènes.
8	non confirmée	Oui	Non	Non	La définition qui a été ajoutée automatiquement n'est pas confirmée par une annalyse automatique des textes
4	invalidée manuellement	Non	Non	Non	L'opérateur à choisit d'invalider la définition.
6	désactivée sur liste	Non	Non	Non	La définition concerne un label jugés abérant (correspond au réimport)
7	transférée	Non	Non	Non	la définition a été transféré du gène fils au gène père. Elle n'est plus active chez le fils.

Les valeurs des champs *valide*, *prendre* et *confirmation* indiquent quels traitements doivent avoir lieu sur les définitions concernées. L'utilisation de chacun de ces champs est expliquée dans les sections qui suivent.

1er) Structure de données pour permettre la mise à jour

La possibilité de mettre à jour le dictionnaire des gènes est assurée par le champ *Valide*. Les enregistrements qui ne sont pas conformes n'ont pas été supprimés. A la place de cela, nous les avons marqués par le champ *valide*. De cette façon, on garantit qu'ils ne seront pas réintégrés de nouveau dans les données lors des futures mises à jour.

2e) Structure de donné pour prendre ou ne pas prendre en compte les définitions

Certaines définitions ne doivent pas être prises en compte dans l'identification des gènes, même si elles sont exactes. Il s'agit par exemple, des définitions imprécises qui ont été définies dans la section Partie 2Chapitre 1I.D.1. Le champ *prendre* sert à indiquer que les définitions associées ne doivent pas être prises en compte lors de l'identification des gènes. La valeur par défaut pour ce champ est *Oui*.

3e) Structure de données pour exiger la confirmation de la reconnaissance d'une définition

Certaines définitions n'ont pas les qualités suffisantes pour qu'elles puissent être interprétées dans un texte à elles seules. L'identification du gène qu'elles définissent doit être confirmée par ailleurs dans le résumé par la présence d'au moins une autre définition du même gène. Ces définitions sont dites ***à confirmer***. Cette notion à été abordée à la section Partie 2Chapitre 1I.F.1. Le champ *à confirmer* est utilisé dans le processus d'identification des gènes pour vérifier si la reconnaissance de la définition, même isolée, peut être utilisée dans l'interprétation.

4. Structure de données pour l'identification des gènes

L'identification des gènes se fait en deux temps. Dans un premier temps, les textes sont parcourus à la recherche des labels. Le résultat de cette opération est mis dans la table de *reconnaissance des labels*. Dans un second temps, à partir de cette table et en utilisant le dictionnaire des gènes, ces labels sont interprétés, c'est-à-dire associés à un gène. Il s'agit de la reconnaissance des définitions de gènes.

a. Structure de données pour la reconnaissance des labels

Le résultat de la reconnaissance des labels est contenu dans la table dont le nom est *RNG* pour *reconnaissance nom de gènes*. Le tableau 57 donne un exemple d'enregistrement de cette table.

Tableau 57 Table *Reconnaissance des labels*
Ce tableau présente un enregistrement de la table.

Champ	Contenu
Clef Index	70699
N° Phrase	21097
Label	ptc
Position	39
Date	14/09/00 14:15:23

Cet enregistrement correspond à la phrase qui est donnée dans l'exemple 4. La position correspond au rang du premier caractère qui compose le label. La méthode d'indexation des textes sera expliquée dans la section Partie 2Chapitre 2I.B.2.

b. Structure de données pour la reconnaissance des définitions

i. La table des reconnaissances des définitions

Après avoir reconnu les labels, ce sont les définitions de gènes qui sont reconnues. L'information est placée dans la table reconnaissance des définitions dont nous donnons un

enregistrement dans le tableau 58. La table est intitulée *RDG* pour *Reconnaissance Définition de Gène*.

Tableau 58 Table reconnaissance des définitions
Le tableau donne un exemple d'enregistrement de la table.

Champ	Contenu
Clef Reconnaissance	1639
N° Phrase	21097
Position	39
N°Définition	19078
Problème	
Processus	Expert
Redondant	Non
Date	21/04/00 17:58:38

L'exemple que nous donnons là correspond à la première occurrence du gène *patched (pct)* dans l'exemple 4. La définition du gène est celle qui associe le label *pct* au gène *patched (pct)*.

Le champ *redondant* permet de noter que la reconnaissance du gène a déjà eu lieu avec le mot qui précède. Ce concept est expliqué dans la section Chapitre 1I.F.3.

La signification du champ *Processus* est donnée dans la section qui suit.

ii. Structure de données pour savoir quel est le processus d'indexation qui a été mis en œuvre

L'inventaire des modalités possibles du champ *processus* de la table des reconnaissances est réalisé grâce à la table *processus* dont le tableau 59 donne le contenu. La table est intitulée *PRG* pour *Processus Reconnaissance Gène*.

Tableau 59 Table des processus
La table donne la nature du processus qui a conduit à la reconnaissance d'une définition de gène.

Clef	Processus	Prendre	Auto	Commentaires
1	non renseigné	Non	Oui	Pas de valeur.
2	Expert	Oui	Non	Reconnaissance faite par l'annotateur.
3	Définition ignorée	Non	Oui	La définition ne doit pas être prise en compte dans la reconnaissance.
4	Reconnaissance confirmée	Oui	Oui	Reconnaissance confirmée. Une autre définition du même gène est reconnue dans les textes.
5	Label non confirmée	Non	Oui	Défault de confirmation du label.
6	Définition non confirmée	Non	Oui	Défault de confirmation de la définition.
7	Reconnaissance multiple isolée	Non	Oui	Défault de confirmation de la reconnaissance multiple.
8	Reconnaissance simple	Oui	Oui	Reconnaissance simple non confirmée.

Les *processus* sont généralement automatiques (champ *auto*) sauf dans un processus qui est le processus *expert*. Les reconnaissances sont considérées comme à prendre (champ *prendre*) ou à laisser selon le processus qui leur a donné naissance. Les *processus* seront expliqués dans la partie B.3.

B. MÉTHODE D'IDENTIFICATION DES GÈNES

Dans la section A, nous avons expliqué comment les informations sont représentées à l'intérieur de la base de données. Nous pouvons aborder la façon dont ces informations sont mobilisées pour mener à bien l'identification des gènes.

1. La visualisation et l'exploitation des données dans une base de données relationnelle

Nous avons vu que les données sont réparties à l'intérieur de plusieurs tables même si elles ont à voir les unes avec les autres. Il est cependant possible de les rassembler virtuellement dans des tableaux uniques. Ce sont les requêtes qui vont réaliser cela. Nous verrons aussi comment les requêtes peuvent être lancées successivement à l'aide de macros. Nous présenterons aussi les modules qui permettent d'écrire des programmes.

a. L'utilisation des requêtes

Les ***requêtes*** permettent de rassembler des éléments qui se trouvent dans des tables distinctes. Il s'agit d'une vue sur une partie choisie de l'ensemble des données. A coté des ***requêtes sélection*** qui ne modifie aucune donnée des tables, existent aussi des requêtes *modification*, *ajout* ou *création*. Les requêtes *modification* vont intervenir sur les valeurs de un ou plusieurs champs. Les requêtes *ajout* vont créer de nouveaux enregistrements dans une table existante. Les requêtes *création* donnent naissance à des enregistrements dans une table créée pour l'occasion.

Les requêtes obéissent à un langage standard qui s'appelle le ***SQL*** ou ***Structured Query Language***.

b. L'automatisation des tâches

Les actions répétitives plus complexes sont mémorisées dans des macros et des modules.

i. Les macros

Les ***macros*** permettent de rassembler plusieurs requêtes qui ont avantage à être utilisées ensemble. La macro va lancer chaque requête les unes derrière les autres dans un ordre déterminé.

ii. Les modules

Les ***modules*** permettent d'écrire des programmes dans un langage informatique. Le langage utilisé est le *Visual Basic* qui est une propriété de *Microsoft*. Les modules sont utiles pour automatiser des tâches plus complexe que celles qui sont autorisées par les macros.

2. La détection des occurrences de labels

La détection des occurrences de labels consiste à remplir la table *reconnaissance des labels* dont la structure est décrite dans la section A.4.a. Cette opération est réalisée par la macro *indexer*. Elle est faite en deux étapes principales : l'indexation simple puis l'épuration de l'index. Pour plus de commodité, nous utiliserons le terme d'index pour désigner la table de *reconnaissance des labels*.

La reconnaissance des labels est commandée par la macro *indexer*. Les types de traitement à faire sur chaque label sont indiqués grâce à la table *type de reconnaissance* qui est décrite dans la section A.3.b.v.

a. Indexation des textes

L'indexation consiste à parcourir les textes à la recherche des termes contenus dans la table des labels. Cela est fait au niveau des phrases et non au niveau des résumés. L'algorithme utilisé est le plus simple et n'est donc pas le plus rapide. Il consiste à prendre chaque texte l'un après l'autre et à rechercher une chaîne de caractères à l'intérieur de celui-ci. Quand celle-ci est trouvée, le système vérifie que le caractère qui précède, s'il existe, fait bien partie d'une liste prédéfinie de caractères séparateurs. Il en est de même pour le caractère qui suit la chaîne de caractères. Si une occurrence du label est trouvée, l'information est consignée directement dans la table de reconnaissance des labels. Le parcours du texte reprend alors là où il en était rendu. Quand un texte est entièrement parcouru à la recherche d'un label, on passe au label suivant. Quand la liste des labels est épuisée, on passe au texte suivant, jusqu'à épuisement des textes.

Le module chargé de réaliser l'opération est nommé *indexation*.

Le module est utilisé dans la macro *indexer*.

La macro et le module utilisent des requêtes qui indiquent quels sont les textes à indexer, quels sont les syntagmes à rechercher dans les textes et où doit être placé le résultat. Seuls les labels de type de reconnaissance *indexer* sont pris en compte.

L'indexation de 500 résumés prend environ deux heures. Les étapes suivantes ont des temps de calcul négligeable relativement à l'indexation.

b. Correction pour les mots ambigus en début de phrase

Les labels qui sont ambigus quand ils sont placés en début de phrase, reçoivent un traitement particulier. Cela a été motivé dans la section Chapitre 1I.C.1 et les exemples sont donnés dans le tableau 15. Les labels concernés sont repérés grâce au champ *désindexer si début* de la table *type de reconnaissance*. Les reconnaissances associées sont supprimées de l'index dans les cas où elles ont lieux en première position de la phrase.

c. Épuration de l'index

L'épuration des textes est rendue nécessaire par le phénomène d'inclusion des labels qui a été décrite dans la section Chapitre 1I.B.5. Dans l'exemple qui est donné (Exemple 6 Inclusion des labels), on voit que l'occurrence soulignée du label *Hairless* ne doit pas être reconnue. En effet, elle est incluse dans l'occurrence du label *Suppressor of Hairless*. L'épuration va consister à la supprimer de l'index. Elle a lieu quand quatre conditions se trouvent réunies.

- Le label à supprimer éventuellement, par exemple *Hairless*, se trouve en position P_0 dans le texte
- Ce label est inclus dans un autre label, par exemple *Suppressor of Hairless*, en position P_1
- Ce deuxième label est présent dans le texte en position P_2
- Avec la relation $P_0 = P_2 + P_1 - 1$

Pour le point 2, on voit que l'on a besoin de la table d'inclusion dont la structure est décrite dans la section A.3.b.ii. Cette épuration se fait grâce à deux requêtes qui sont incluses dans la macro *indexer*.

d. Reconnaissance des mots vides

Les *mots vides* ne sont pas recherchés à la première passe car cela ferait exploser la taille de l'index. Un ***mot vide*** n'est recherché dans une phrase que s'il participe à la définition d'un gène dont un nom au moins a été trouvé dans la même phrase. Les labels concernés sont repérés grâce au champ *2ième vague* de la table *type de reconnaissance*.

3. Interprétation des labels

Nous avons vu qu'un même terme peut être interprété de plusieurs façons, soit qu'il puisse représenter autre chose qu'un gène, soit qu'il puisse renvoyer vers plusieurs gènes. Il s'agit donc bien d'interpréter la présence d'un label. Ainsi, la table des reconnaissances de label doit être retravaillée pour servir à compléter la table de reconnaissance des définitions. La macro qui réalise cette opération se nomme *interprétation*.

L'interprétation se fait en plusieurs étapes dont nous donnons maintenant le détail. Chaque étape correspond à une valeur dans le champ *processus* de la table *reconnaissance des définitions*. Cette table est donnée dans le tableau 59.

La première étape consiste à aller compléter tout simplement la table des reconnaissances à partir des informations contenues dans l'index et dans le dictionnaire des gènes. A ce stade, la valeur du processus est *non-renseigné*.

Les étapes suivantes consistent à effectuer des tests successifs sur les enregistrements ajoutés précédemment à la table des reconnaissances et à modifier la valeur du champ *processus* si le test s'est avéré positif. A la fin des opérations, plus aucun enregistrement n'a un processus qui est encore sur la valeur *non renseigné*.

Voyons maintenant la signification de chacun de ces processus de reconnaissance.

Le processus *définition ignorée* correspond à des définitions qui ne doivent pas être prises en compte. Cette information est contenue dans la table *confiance dans la définition* dont la structure a été décrite dans la section A.3.c.iii : le champ *prendre* des enregistrements concernés a la valeur *non*.

Le processus *reconnaissance confirmée* correspond à l'identification d'un gène qui a déjà été reconnu par ailleurs dans le résumé à l'aide d'une autre définition.

Le processus *label non-confirmé* correspond aux reconnaissances de labels qui auraient dû être confirmées, mais qui ne le sont pas. L'information de cette exigence posée sur le label est contenue dans la table *type de reconnaissance* dont la structure a été décrite dans la section A.3.b.v.

Le processus *définition non-confirmée* correspond aux définitions qui auraient dû être confirmées, mais qui ne le sont pas. Cette exigence sur la définition est inscrite dans la table *confiance dans la définition* qui a été décrite dans la section A.3.c.iii.

Le processus *reconnaissance multiple isolée* correspond à des reconnaissances non confirmées qui sont cependant multiples. La notion de reconnaissance multiple a été introduite dans la section Chapitre 1I.F.1.

Une fois tous ces tests successifs effectués, on est nécessairement dans le cas d'une reconnaissance simple et non confirmée qui est le dernier des processus.

Une fois que le processus a été complété, il reste le cas de la reconnaissance redondante. Le test ne s'effectue bien sûr que sur les reconnaissances qui ont déjà été caractérisées comme confirmées. Il consiste à rechercher si la reconnaissance confirmée n'est pas précédée immédiatement avant, en terme de position dans le texte, par une reconnaissance du même gène. La notion de reconnaissance redondante a été introduite dans la section Chapitre 1I.F.3

On remarquera qu'à chaque processus est associée une décision : *prendre* ou *ne pas prendre*.

Un exemple de résumé interprété est donné dans la section Chapitre 3I.B.2.

L'interprétation de 500 résumés se fait en environ 10 minutes. C'est donc beaucoup plus rapide que l'indexation.

C. ACQUISITION DES DONNÉES NÉCESSAIRES À L'ANALYSE

1. Collecte des textes et intégration dans la base de données

a. Choix des résumés Medline

La constitution de l'échantillon s'est faite par rapport au fichier de phrases issu de *Flybase* qu'a étudié PILLET. Ces phrases sont des notes prises par les annotateurs de *Flybase* à partir de publications pour la plupart référencées par *Medline*. Nous avons décidé de prendre comme échantillon les résumés *Medline* de ces publications. Nous avons justifié ce choix dans la section Partie 1Chapitre 3II.A.2. Techniquement nous avons suivi le lien qui existe entre les phrases du fichier que nous a donné PILLET et des enregistrements d'une base de données de références bibliographiques issues de *Flybase*. Ces références bibliographiques citent elles-mêmes des références *Medline*. En conjuguant les deux liens, nous obtenons un lien direct entre les données qu'a étudiées PILLET et les données que nous avons étudiées. Ce lien est représenté par une référence au résumé dans la table des phrases issues de *Flybase* que nous présentons dans le tableau 60.

Tableau 60 Table phrase Flybase
Ce tableau présente un enregistrement de la table.

Champ	Contenu
Clef	3597
Symbole	#su(w[a])
N° Flybase	46126
Prase	Control of su(w[a]) is controlled at the level of splicing and this control represents autorepression of su(w[a]) expression
Phrase1	CONTROL #su(w[a]) CONTROL LEVEL SPLICE CONTROL REPRESENT AUTOREPRESS #su(w[a]) EXPRESS
Phrase2	CONTROL CONTROL LEVEL SPLICE CONTROL REPRESENT AUTOREPRESS EXPRESS
Validation	Y
Direction	Y
Type	R
Interaction	su(w[a])_su(w[a])/R
Problème	
N° Medline	88166655
Titre	Evidence that a regulatory gene autoregulates splicing of its transcript.
Auteurs	Zachar Z^Chou TB^Bingham PM
IVI	0,095

Les données fournies par PILLET ont été intégrées telles quelles, c'est à dire que nous n'avons pas cherché à les structurer selon une logique de base de données relationnelles. La technique d'import sera décrite dans la section b.iii.

Certaines phrases issues de *Flybase* citent des résumés qui sont absents de notre base de données. Il s'agit de résumés qui étaient absents de l'ensemble des résumés que nous avons extraits de *Medline*.

Une partie des résumés de notre base de données ne sont cités par aucune phrase *Flybase*. Il s'agit des résumés d'origine *sans nom de gène*. Nous avons présenté la table *origine* dans la section A.2.a.ii.2e).

b. Intégration des textes issus de Flybase et de Medline

i. Import des textes issus de Medline

L'import des résumés s'est fait à partir de dix-huit cédéroms qui couvrent la vie de *Medline* depuis sa création en 1966 jusqu'en avril 1998. Sur chaque cédérom, nous avons fait une requête qui consiste à demander les références qui citent au moins une fois la drosophile. Cette interrogation est faite sur le champ *résumé,* et nous avons utilisé une *troncature gauche* pour ne manquer aucun résumé. Les fichiers correspondants à chaque interrogation ont été concaténés puis importés dans le gestionnaire de base de données. Nous avons ensuite sélectionné le sous-ensemble des résumés qui correspondent aux mêmes publications que celles dont il est question dans le corpus de PILLET. Ceci en conformité avec le choix de l'échantillon d'analyse que nous avons fait et que nous avons expliqué dans la section Partie 1Chapitre 3II.A.2. Cela a réduit la base de 19410 à 529 résumés. Dans un deuxième temps, nous avons interrogé le premier corpus des 19410 résumés pour rechercher les résumés qui ne comportent aucun nom de gènes. Pour ce faire, nous avons exclu de la liste des noms de gènes les *mots vides* les plus répandus de façon à obtenir un nombre de résumés sans nom de gène de taille suffisante. Les 215 résumés ainsi obtenus ont été importés dans la base de données.

Access n'admettant comme format d'import que les textes sous forme de tableaux, nous avons utilisé un logiciel de reformatage de fichiers texte du nom d'*Infotrans*.

ii. Éclatement des résumés en phrases

Les résumés fournis par *Medline* sont composés de phrases. Le passage d'une phrase à l'autre a été marqué par la chaîne de caractère ^. (un accent circonflexe puis un point). Ce formatage réalisé par *Medline* grâce à des méthodes automatiques comporte des erreurs mais en nombre assez faible. Nous n'avons pas cherché à corriger ces erreurs. Par ailleurs, pour des raisons purement techniques, nous avons transformé ce séparateur basé sur deux caractères en un séparateur basé sur un seul. Il s'est agit de | (la barre verticale), qui était par ailleurs absente des textes.

Les résumés ont été segmentés (ou éclatés) en phrases à l'aide d'un module que nous avons conçu pour réaliser ce type d'opération et qui se nomme *Eclater*. La méthode consiste à parcourir le texte des résumés à la recherche d'un caractère spécifique. Ici le caractère spécifique est la barre verticale. Quand il est rencontré, un nouvel enregistrement est créé dans la table des phrases. Le texte de la phrase est alors complété à l'aide du segment de texte du résumé qui précède le séparateur. La référence au résumé est aussi placée dans l'enregistrement *phrase*, ainsi que le rang de la phrase dans le résumé.

Le module segmenté est commandé par la macro *éclater les résumés*.

iii. Import des textes issus de Flybase

Les phrases de l'étude de PILLET nous ont été fournies sous forme d'un fichier texte. Nous avons importé ces données dans une table unique. Après dédoublonage nous obtenons 929 phrases.

2. Constitution des données relatives au dictionnaire des gènes

a. Importation des données terminologiques

Les données utilisées proviennent d'un export de *Flybase* en format texte. Ce fichier de 50 Méga Octet a d'abord été filtré pour exclure toutes les données relatives aux allèles. Le fichier a ensuite été reformaté à l'aide d'*Infotrans*. Le codage des lettres grecques a été refait pour le mettre en conformité avec le codage de *Medline*. Les données relatives aux ***chimères***, qui sont des assemblages de plusieurs gènes par génie génétique, ont aussi été exclues automatiquement. Ces diverses opérations aboutissent à la création de plusieurs fichiers de textes au format tableau qui sont importables dans le logiciel de gestion de base de données.

b. Les étapes de filtrages et de reformatages

Une fois les données importées dans le logiciel de gestion de base de données, de nouveaux filtrages ont encore été nécessaires. Certains *noms synonymes* étaient formatés sous la forme *nom abrégé : nom complet*. Il s'est agi de transformer cela de façon à ne mettre qu'un seul nom par enregistrement. Des commentaires comme *Unammed* (non nommé) ou le point d'interrogation ont aussi dû être supprimés. D'autres commentaires placés entre parenthèses qui indiquaient que le terme proposé était déjà pris pour un autre gène ont aussi été supprimés sachant que nous avons cette information par ailleurs.

c. Mise en forme relationnelle

Les données importées ont été retraitées pour se conformer au schéma relationnel de la base. En effet, après l'import, les données sont présentes dans une table unique. Ainsi des informations hétérogènes se retrouvent mêlées dans une structure unique. Par exemple, une table fraîchement importée va pouvoir donner simultanément des informations sur les gènes comme le numéro *Flybase* de chaque gène, et des informations sur des définitions de gènes. Pourtant ces deux informations doivent être placées dans des tables distinctes d'après le schéma relationnel de la base tel que nous l'avons défini. Un travail de structuration des données est nécessaire pour mettre les données en conformité avec la structure que nous avons choisie.

Le fichier qui sert à l'import est dit ***plat***. Cela signifie que dans ce fichier, les données hétérogènes, que sont labels, gènes, définitions etc., ne sont pas nettement séparées. A l'opposé, dans un fichier de base de données relationnelle, les données hétérogènes sont clairement séparées car elles sont placées dans des tables distinctes.

La mise en forme relationnelle consiste à aller extraire de la table importée les informations qui doivent être placées dans les différentes tables. Ceci est réalisé par une succession de requêtes. Cette opération n'a été faite qu'une seule fois. Nous n'avons donc pas créé de procédure pour automatiser cette tâche comme cela a été fait pour l'identification des gènes.

d. Préparation de l'indexation des textes

Avant de pouvoir indexer les textes, il est nécessaire de disposer de la table d'inclusion dont la structure est décrite dans la section A.3.b.ii. Cette table a été obtenue par l'utilisation du module d'indexation.

La macro *reconstruire RNGR* construit la table après avoir supprimé tous les enregistrements de celle-ci. Quatre nuits sont nécessaires à cette construction.

La macro *actualiser RNGR* permet la mise à jour de la table. Les enregistrements concernés par cette mise à jour sont repérés par le champ *nouveau* de la table des labels. Le principe est le suivant :

La valeur du champ nouveau est par défaut *oui*. Chaque enregistrement créé dans la table des labels sera donc considéré comme nouveau. Il en sera ainsi jusqu'à la mise à jour de la table d'inclusion. En effet, à la fin du processus de mise à jour, tous les champs de la table des labels seront positionnés sur *non-nouveau*. L'actualisation des données de *RNGR* se fait de la manière suivante :

L'indexation a lieu comme si toute la table était à reconstruire sauf que tous les labels ne sont pas indexés et que tous les labels eux-mêmes ne servent pas à l'indexation. Dans un premier temps, seuls les labels nouveaux sont indexés. Dans un deuxième temps, tous les labels sont indexés mais seuls les labels nouveaux servent à l'indexation.

e. Complémentation du dictionnaire

i. Ajout de nouvelles entités biologiques qui ne sont pas des gènes

Lors de l'annotation des textes, nous nous sommes rendu compte qu'il était nécessaire d'ajouter des rubriques au dictionnaire de gènes pour y inclure des entités biologiques qui ne sont pas des gènes mais qui dans les textes jouent le même rôle. Nous avons déjà évoqué cette question dans la section Chapitre 1I.B.5.b.

Ces ajouts ont été faits manuellement par les annotateurs au fur et à mesure des besoins. Pour maintenir la cohérence du dictionnaire, il a été nécessaire de transférer des définitions. Ainsi par exemple, lors de l'introduction de la famille de protéines *actin*, il a fallu transférer les définitions relatives aux gènes *Actin 42A (Act42A)*, *Actin 57B (Act57B)*, *Actin 5C (Act5C)*, *Actin 79B (Act79B)*, *Actin 87E (Act87E)* et *Actin 88F (Act88F)*. Le transfert est matérialisé grâce à la modalité *transférée* du champ *confiance* des définitions concernées. Le transfert implique bien une invalidation de la définition comme on peut le voir dans le champ *Valide* de la table *confiance dans les définitions*. La destination du transfert est présente à travers un lien indirect. En effet, les deux enregistrements de la table *gène ou assimilé* sont mis en relation par l'intermédiaire de la table *filiation*.

ii. Ajout de termes spécifiques

Lors de l'annotation des textes, nous avons constaté que des confusions étaient possibles entre les noms des gènes et d'autres termes du domaine de la génétique. Nous avons déjà évoqué cette question dans la section Chapitre 1I.B.5.b. Ces termes spécifiques ont été rajoutés à la main et au fur et à mesure des besoins. Ils ont été rajoutés à la table des labels de façon à être utilisés lors de l'indexation des textes. Cependant leur origine étrangère a été marquée grâce à la modalité *terme spécifique* du champ *reconnaissance* de la table des labels.

iii. Caractérisation de l'ambiguïté des labels

Lors de l'annotation des textes, nous nous sommes rendu compte que certains labels étaient ambigus, de sorte qu'un traitement particulier devait leur être réservé dans le processus d'identification automatique des gènes. Cette difficulté a été évoquée dans la section Chapitre 1I.C. Il a donc fallu faire un inventaire des termes ambigus. Cela a conduit à la caractérisation du traitement à faire sur chaque label dans le champ *type de reconnaissance* de la table des labels.

Cette caractérisation du degré d'ambiguïté de chaque label a été faite de diverses manières. Nous avons tout d'abord bénéficié de l'expérience de PILLET qui nous a transmis des listes de noms de gènes ambigus. Nous avons ensuite fait diverses statistiques sur la présence des labels dans les textes. L'une d'entre elles est expliquée dans la section 3.c. D'une manière générale ces statistiques avaient pour but de détecter des anomalies dans les fréquences des gènes.

Enfin nous avons aussi corrigé les valeurs d'ambiguïté des labels au cours de l'annotation au vu des difficultés qui se présentaient.

3. Acquisition de nouvelles connaissances sur la nomenclature des gènes

a. Construction des définitions variantes

Nous avons vu dans la section Chapitre 1I.D.2 que le dictionnaire de *Flybase* ne prévoit pas toutes les orthographes possibles pour chaque gène. Nous avons donc introduit de nouvelles définitions inspirées de celles qui se trouvaient déjà dans le dictionnaire. Ces définitions variantes ont été construites automatiquement. Nous avons tout d'abord fait l'inventaire des variations orthographiques pertinentes sur les labels. Ensuite pour chaque type de variation, nous avons généré les définitions correspondantes. Dans certains cas cela amène à réutiliser un label qui est déjà pris par ailleurs. Par exemple, la définition *dl* du gène *dorsal (dl)* va être transformée dans une définition de *dorsal (dl)* par *Dl* alors que ce label est déjà pris pour le gène *Delta (Dl)*. Ces définitions sont donc supprimées. Dans un deuxième temps nous avons complété la table des labels avec les variations introduites. Puis nous avons rétabli le lien entre la table des définitions et la table des labels. Dans un troisième temps nous avons mis à jour la table des transformations.

Toutes ces opérations se font à l'aide de requêtes. L'enchaînement de ces requêtes permet d'obtenir le résultat escompté.

b. Validation des définitions par l'analyse des textes

La validation des définitions consiste à vérifier si la définition est crédible ou ne l'est pas. Elle est faite automatiquement et par l'analyse des textes. Cette validation des définitions a été faite uniquement sur les définitions *variantes*. Elle pourrait cependant être généralisée avec quelques adaptations.

Les définitions *variante*s ne sont pas utilisables telle quelles. En effet, il arrive fréquemment qu'elles introduisent de nouvelles difficultés, car les labels obtenus par transformation des labels existants ont toutes les chances de désigner autre chose que le gène qu'ils sont censés définir.

Il s'agit donc de faire le tri entre les bonnes et les mauvaises définitions variantes. Ceci est rendu possible par l'utilisation du contexte, comme nous l'avons présenté dans la section Chapitre 1I.F.4.

Cette opération se fait après avoir lancé le processus d'identification automatique des gènes dans les textes. Les définitions variantes, qui ne sont jamais confirmées dans aucun résumé, sont alors caractérisées de *non confirmé*. Cette information est consignée dans le champ *confiance* de la table des définitions. Ceci est fait à l'aide de la requête intitulé *CD=non confirmé*.

c. Validation des labels par l'analyse des textes

L'analyse des textes permet aussi de valider le caractère univoque des labels. Le principe de cette analyse consiste à observer des anomalies dans les statistiques de fréquence des labels. Diverses statistiques ont été utilisées, mais l'imbrication entre des phénomènes, tels que l'ambiguïté et l'imprécision de la terminologie, fait qu'il n'est pas possible de trouver un indicateur statistique qui permette à coup sûr de faire la différence entre un label ambigu et un label qui ne l'est pas. Ainsi, chaque méthode a ses inconvénients et aucune ne peut être utilisée en aveugle. Cependant chacune permet de pointer sur des labels potentiellement

ambigus et il appartient à l'opérateur de décider de la caractérisation des labels ainsi désignés.

La méthode la plus simple consiste à classer les labels par ordre de fréquence décroissante comme dans le tableau 61.

Tableau 61 Ambiguïté et fréquence
Le tableau donne les labels les plus répandus. La plus part d'entre eux sont ambigus. La fréquence désigne ici le nombre d'occurrence du terme dans les textes.

Label	Fréquence
in	362
cell	275
to	205
is	177
dpp	133
early	116
Ubx	97
similar	96
Notch	89
Sxl	79
wingless	78
as	73
bcd	72
D	69
eye	68
dorsal	65

On voit que dans ce cas, la difficulté provient du fait que certains gènes comme *Ultrabithorax (Ubx)* sont si répandus dans les textes que leurs labels, tel *Ubx*, viennent côtoyer les termes très ambigus comme *similar*.

II. MISE EN ŒUVRE DE LA RECONNAISSANCE AUTOMATIQUE DES INTERACTIONS

A. STRUCTURE DE DONNÉES POUR LA RECONNAISSANCE DES INTERACTIONS

1. Table de reconnaissance des interactions

Les annotations sur les textes qui concernent les interactions sont consignées dans la table *Reconnaissance des interactions*. Cette table a pour nom *CI* pour *reconnaissance des Couples de gènes en Interaction*. Le tableau 62 donne un exemple d'enregistrement.

Tableau 62 Table reconnaissance des interactions
Ce tableau donne un exemple d'enregistrement de la table. Dans la phrase 22180, l'expert a reconnu une interaction entre *achaete (ac)* et *hairy (h)*, cette interaction étant orientée de second vers le premier.

Champ	Contenu
Clef Interaction	816
N° Phrase	22180
Gène 1	achaete (ac)
Gène 2	hairy (h)
Problème	
Ordre	Sens inverse
Gène 1 absent	Non
Gène 2 absent	Non
Processus	Expert
Date	21/04/00 17:08:15

Par convention, le gène 1 est celui qui a le plus petit numéro et le gène 2 est celui qui a le plus grand numéro. La direction est précisée dans le champ *Ordre* dont nous verrons les modalités dans la section 2.

Les modalités du champ *Processus* seront expliquées dans la section 3.

Les champs *Gène 1 absent* et *Gène 2 absent* servent à valider l'information selon laquelle le partenaire numéro 1 ou 2 est absent de la phrase concernée. En effet, cette information est en principe contenue dans la table de reconnaissance des définitions de gènes, mais dans certains cas, l'annotateur peut juger que d'une certaine façon le gène est présent même si aucune de ses définitions n'a été reconnue. Ceci est illustré dans l'exemple suivant.

Exemple 24 Partenaire présent mais non reconnu
Dans cette phrase une interaction entre *dorsal (dl)* et *cactus (cact)* est implicitement affirmée par la présence du complexe *dl2cact*, cependant le gène *dorsal (dl)* n'est pas reconnu car *dl2cact* n'est pas considéré comme un label de *dorsal (dl)*. La présence du champ *Gène absent 2* permet à l'annotateur de préciser que *dorsal (dl)* n'est pas absent de la phrase contrairement aux apparences.

Complex 2 (270 kDa) consists of one complex 1 and one cact molecule (dl2cact).

2. Table Ordre dans les interactions

La table *Ordre dans les interactions* (*OI*) réalise l'inventaire des modalités possibles dans le champ *ordre* de la table R*econnaissance des interactions*. Le tableau 63 donne le contenu de cette table.

Tableau 63 Table Ordre des interactions
La table fait l'inventaire des modalités possible du champ Ordre de la table des reconnaissances d'interactions.

Clef	Ordre	Commentaires
1	Indéterminé	Pas d'informations sur le sens de l'interaction.
2	Sens direct	Le gène 1 modifie l'expression du gène 2.
3	Sens inverse	Le gène 2 modifie l'expression du gène 1.
4	Double sens	Le gène 1 modifie l'expression du gène 2 et réciproquement.

La présence d'une interaction clairement réciproque dans une même phrase est très rare. Nous avons pu le constater lors de l'annotation de l'échantillon A. Ainsi, la modalité *double sens* est très rare. Cependant, nous avons préféré considérer une interaction *double sens* comme un type particulier d'interaction plutôt que de considérer qu'une même phrase pouvait contenir plus d'une interaction concernant les mêmes partenaires.

3. Table Processus de reconnaissance des interactions

L'annotation des phrases, en ce qui concerne les interactions, est réalisée soit par le programme, soit par un expert. Dans la table *Reconnaissance des interactions*, c'est le champ *Processus* qui permet de préciser cela. Les modalités possibles pour le champ sont consignées dans la table *PRI*—pour *Processus de reconnaissance des interactions*—. Le tableau 64 donne le contenu de cette table.

Tableau 64 Table Processus de reconnaissance des interactions
Cette table réalise l'inventaire des modalités possible du champ Processus de la table de reconnaissance des interactions.

Clef	Processus	Commentaires
1	Expert	Interaction écrite par l'annotateur.
2	*2RDG*	Interaction écrite par l'ordinateur dans les phrase où deux définitions de gènes ont été reconnues.
3	*nRDG*	Interaction écrite par l'ordinateur dans les phrases où plusieurs définition de gène ont étés reconnues.
4	*2G*	Interaction écrite par l'ordinateur dans les phrases où deux gènes ont été identifiés.
5	*nG*	Interaction écrite par l'ordinateur dans les phrases ou plusieurs gènes ont été identifiés.

B. STRUCTURE DE DONNÉES POUR L'IVI

1. Structure de données pour le dictionnaire de lemmatisation

Les statistiques calculées sur les textes utilisent des donnés qui sont relatives à des lemmes et non à des formes fléchies. Un dictionnaire de lemmatisation est donc inclus dans la base de données. Le repérage du vocabulaire spécifique se fait sans la prise en compte de la différence entre les majuscules et les minuscules. Les données du dictionnaire sont, par convention, toutes en minuscules.

a. Structure de données pour les lemmes

Les lemmes sont contenus dans la table des *lemmes* (Lm) dont le tableau 65 donne un exemple d'enregistrement. La table compte 436 enregistrements.

Tableau 65 Table des lemmes
Le tableau suivant donne un exemple d'enregistrement de la table. Les facteurs sont issus de l'analyse factorielle de PILLET. Seul les cinq premiers champs ont été utilisés.

Champ	Contenu
N° Lemme	6149
Lemme	downstream
Fréquence Oui	15
Fréquence Non	0
Fréquence Indéterminé	0
Facteur principal	-1,02378305
Facteur secondaire	-0,07715504
Catégorie	y
Pouvoir Discriminant	2,589047

b. Structure de données pour les formes fléchies

Les formes fléchies sont contenues dans la table des *mots bruts* (MB) dont le tableau 66 donne un exemple d'enregistrement. La table compte plus de 2000 enregistrements.

Tableau 66 Table des formes fléchies
Le tableau donne un exemple d'enregistrement. Le champ lemme fait référence à un enregistrement de la table des lemmes.

Champ	Contenu
N° Mot brut	436
Mot brut	activated
Lemme	activate
Date	21/04/00 17:53:22

2. Structure de données pour la reconnaissance des formes fléchies

La méthode des *IVI* est basée sur la présence d'un vocabulaire spécifique dans les phrases. Les flexions étant prises en compte, se sont les formes fléchies qui sont recherchées dans les textes. L'index correspondant est contenu dans la table *RMB* (*Reconnaissance Mots Bruts*) dont le tableau 67 donne un exemple d'enregistrement.

Tableau 67 Table de reconnaissance des formes fléchies
Le tableau donne un exemple d'enregistrement de la table. Le terme *requires* à été reconnu en position 74 dans la phrase de numéro 17406.

Champ	Contenu
N° Phrase	17406
Mot brut	requires
Position	74
Date	21/04/00 17:13:05

C. CONSTITUTION DES DONNÉES RELATIVES AU DICTIONNAIRE DE LEMMATISATION

Les données relatives au dictionnaire de lemmatisation ont été fournies par PILLET. Le dictionnaire est adapté au travail sur des textes de génétique, car il a été complété manuellement par PILLET pour lemmatiser intégralement les textes issus de *Flybase* qui font partie de son échantillon d'étude.

Les données ont été importées à partir d'un fichier texte formaté comme ceci : "forme fléchie en minuscule","forme lemmatisée en majuscule". Les données ont été dédoublées de façon à ce qu'une forme fléchie ne renvoie qu'à une seule forme lemmatisée. Cette opération est un préalable à la mise en forme relationnelle. Pour des raisons d'encombrement, seule la partie du dictionnaire qui concerne les termes spécifiques de l'interaction a été conservée dans la base de données.

D. MÉTHODE DE RECONNAISSANCE DES INTERACTIONS

1. Calcul de l'*IVI*

La spécificité d'un lemme a été définie dans la section Partie 1Chapitre 3I.C.1 comme la proportion des textes qui décrivent une interaction parmi celles qui utilisent le lemme. La spécificité d'une forme fléchie est définie comme la spécificité de la forme lemmatisé associée.

L'*IVI* d'une phrase est défini (cf. Définition 2) par la moyenne des spécificités des termes reconnue. La requête permettant le calcul est donnée dans l'exemple 25.

Exemple 25 Requête SQL de calcul des IVI
Cette requête permet le calcul de l'IVI de chaque phrase.

SELECT RMB.IPM, Avg(Y/(Y+N+I)) AS IVI

FROM (Lm INNER JOIN MB ON Lm.ILm = MB.ILm) INNER JOIN RMB ON MB.IMB = RMB.IMB

GROUP BY RMB.IPM;

Cette requête peut se traduire par :

A partir du tableau construit en considérant les enregistrements de la table des lemmes (Lm) qui sont cités par la table des formes fléchies (MB) et qui eux même sont cités par la table des reconnaissances (RMB), en regroupant les enregistrements qui concernent la même phrase, sélectionner le champ numéro de phrase et la moyenne des valeurs Y/(Y+N+I), cette dernière étant désignée par *IVI*.

Le résultat de la requête a été écrit dans la table des phrases. Il n'est pourtant pas nécessaire d'écrire le résultat dans une table. Une solution alternative consiste à faite référence à la requête à chaque fois que l'on a besoin de l'*IVI* d'une phrase. Cependant, nous avons choisit de mettre le résultat dans une table pour que le calcul de l'*IVI* ne se fasse pas à chaque fois. De cette façon, l'affichage des formulaires de consultation des données relatives aux phrases est plus rapide. Par ailleurs, cette structure permet d'utiliser n'importe quel indicateur statistique à la place de l'*IVI*, même si cet indicateur est calculé par un autre programme.

2. Annotation sur les interactions

Des annotations sur les interactions peuvent être crées automatiquement. Elles sont destinées à être comparées aux annotations écrites par l'annotateur des textes.

Nous n'avons pas cherché à créer des reconnaissances d'interaction ordonnées. En d'autres termes, les annotations créées automatiquement sont toutes de type non-ordonnées. De même, nous n'avons créé des annotations que pour les gènes, au sens strict du terme, alors que l'annotateur a pu écrire des annotations qui font intervenir des groupes de gènes comme par exemple la famille de protéines *actin*.

Cette annotation automatique n'a été faite que sur l'*échantillon A*.

Le principe de l'annotation automatique est le suivant : pour chaque couple d'occurrence de gènes, est créé une reconnaissance d'interaction entre les gènes correspondants. La sélection sur le critère des *IVI* n'intervient que dans un deuxième temps.

Le principe d'annotation que nous venons de décrire donne naissance aux annotations de *processus* intitulé ***nRDG***. Les autres processus sont des variantes qui consistent à ne retenir qu'une partie des annotations de ce type selon des critères qui seront précisés dans la partie évaluation dans la section Chapitre 3II.

Nous donnons ci-après un exemple de résumés annotés par le programme.

Tableau 68 Exemple d'annotation automatique d'un résumé
La colonne *interaction possible* donne le résultat du processus d'annotation automatique avant la prise en compte de l'*IVI*. La dernière colonne donne le résultat après prise en compte de l'*IVI*.

Phrase	Interaction possible	IVI	Interaction retenue
The Enhancer of split locus is required during many cell-fate decisions in Drosophila, including the segregation of neural precursors in the embryo		-0.17	
We have generated monoclonal antibodies that recognise some of the basic helix-loop-helix proteins encoded by the Enhancer of split locus and have used them to examine expression of Enhancer of split proteins during neurogenesis		-0.39	
The proteins are expressed in a dynamic pattern in the ventral neurogenic region and are confined to those ectodermal cells that surround a neuroblast in the process of delaminating		-0.18	
There is no staining in the neuroblasts themselves		-0.67	
We have also examined the relationship between Enhancer of split protein accumulation and the Notch signalling pathway	N_E(spl)	0.04	
Protein expression is abolished in a number of neurogenic mutant backgrounds, including Notch, but is increased as a result of expressing a constitutively active Notch product		-0.09	
We conclude that Notch signalling activity is directly responsible for the accumulation of basic helix-loop- helix proteins encoded by the Enhancer of split locus	N_E(spl)	0.13	N_E(spl)

Dans cet exemple, le seuil choisi pour l'*IVI* est de 0,1. Le programme ne fait aucune annotation par erreur et n'omet aucune interaction. Si le seuil avait été choisi inférieur à 0,04 le programme aurait fait une annotation par erreur dans la phrase numéro cinq. Cependant l'interaction extraite de cette phrase aurait été exacte ; il y a bien une interaction entre les deux gènes, comme on peut le voir dans la dernière phrase. On voit bien ici l'intérêt d'un décompte par interaction plutôt que d'un décompte par occurrence d'interaction. C'est ce qui est fait dans le paragraphe qui suit.

A partir des annotations faites phrase par phrase, il est possible de faire un bilan des interactions extraites sur l'ensemble du corpus. Le tableau 69 donne la liste des interactions extraites par le programme sur les phrases qui citent deux gènes.

Le tableau 70 donne la liste des interactions extraites par l'annotateur sur le même ensemble de phrases. Les lignes en gras sont les lignes communes aux deux tableaux. Il y en a 55. Le tableau 69 compte 75 lignes. Le taux de précision est donc de 55/75 soit 73 %. Le tableau 70 compte 62 lignes. Le taux de rappel est donc de 55/62 soit 89 %. Davantage de statistiques seront données dans la partie évaluation.

III. INTERFACE DE VISUALISATION DES DONNÉES CONTENUES DANS LA BASE DE DONNÉES

Nous avons conçu une interface originale de visualisation des données présentes dans la base. Il s'agit de voir sur un même écran plusieurs types d'informations. En particulier nous nous intéressons aux liens entre, d'une part l'*IVI* et les occurrences de gènes et, d'autre part la présence d'interaction. Par ailleurs, nous souhaitons pouvoir confronter annotations automatiques et annotation manuelle.

A. CONFRONTATION ENTRE INDICES ET FAITS SUR LES INTERACTIONS

L'interface que nous proposons permet de visualiser sur des cas concrets la véracité de notre postulat de base. En effet, elle permet pour un résumé donné, de montrer à la fois *IVI*, présence de gènes et présence d'interactions. Le formulaire de consultation des données sur les annotations est présenté sur la figure 2.

Pour cela, trois graphiques sont présentés à gauche de l'écran. Celui du haut permet de connaître l'*IVI* de chaque phrase du résumé. Cela donne tout d'abord une idée globale sur le résumé, à savoir sa propension à décrire des interactions. Cela permet aussi de voir dans quelle partie du résumé il semble qu'il y ait des descriptions d'interaction.

Le graphique du milieu permet de voir quelle sont les gènes en présence dans le résumé. La légende établit la liste de tous les gènes. Les noms utilisés sont ici les *symboles*, de façon à ne pas encombrer inutilement le graphique. Ce graphique permet aussi de connaître la répartition dans les phrases. En particulier on repère immédiatement les phrases qui comptent plus de deux gènes.

On peut alors faire la synthèse des informations contenues dans les deux premiers graphiques en repérant les phrases qui ont un bon *IVI* et qui citent deux gènes au moins. Ce sont les phrases potentiellement porteuses d'interactions.

Le graphique du bas permet de confronter cette prévision avec la réalité. Les interactions y sont représentées.

La partie droite de l'écran donne des informations sur une des phrases du résumé. Dans l'instantané qui est présenté, il s'agit de la phrase 9. La partie supérieure donne le texte de la phrase. La partie médiane donne les identifications de gènes faites dans la phrase. Chaque ligne correspond à une reconnaissance de gènes, avec position, label reconnu et gène attribuée. La partie inférieure donne les interactions reconnues dans la phrase avec partenaires et ordre.

Sur le résumé que nous présentons, nous pouvons faire plusieurs constats, qui nous amènent à des investigations plus précises.

Nous voyons que les choses se sont bien passées sur la phrase 5. Cette phrase à un *IVI* très favorable de 0,08. De plus cette phrase cite un grand nombre de gènes. Elle a donc toutes les chances de décrire des interactions. Nous observons que c'est effectivement le cas. En revanche, il est difficile de dire combien d'interactions seront présentes. Dans cet exemple, il y a 5 gènes dans la phrase, ce qui fait 10 interactions possibles si l'on ne compte pas les auto-interactions et si on ne prend pas en compte la direction des interactions. Dans la réalité, il n'y en a que 7.

La phrase 4 avait un *IVI* bien plus fort mais ne citait qu'un seul gène. On est donc très satisfait de ne voir aucune interaction dans cette phrase. Cependant la présence d'un *IVI* fort parait assez étonnant pour une phrase qui ne décrit pas d'interaction. Une investigation plus poussée nous permet de comprendre le phénomène. En réalité la phrase, qui a déjà été présentée exemple 18 page 81, décrit une interaction, mais cette interaction fait intervenir un complexe de gènes et non un gène. Cette interaction a été notée par l'annotateur mais elle ne figure pas dans le graphique car il ne prend en compte que les gène au sens strict. Il est naturel que l'*IVI* soit grand car il y a bien un vocabulaire de l'interaction. C'est notre

définition restrictive des interactions qui est la source du paradoxe de cette phrase à fort *IVI* mais sans interaction.

Le cas de la phrase 9 est lui aussi favorable. En effet, l'*IVI* de cette phrase est plutôt favorable. Il est de -0,04. Le fait qu'il soit négatif n'est pas déterminant car il n'y à aucune raison de comparer systématiquement les *IVI* à zéro. A titre de comparaison l'*IVI* moyen des phrases liées à *Flybase* est de -0,09. Par ailleurs la phrase cite deux gènes. On peut donc s'attendre à une interaction entre ces deux gènes. Le graphique du bas permet de vérifier que c'est effectivement le cas.

La phrase 8 décrit la même interaction bien que cette fois l'*IVI* ne soit pas favorable. Le système n'a donc pas bien fonctionné avec cette phrase.

En revanche le système a bien fonctionné pour la phrase 6 qui fait intervenir les mêmes gènes mais qui ne décrit pas d'interactions entre eux.

La phrase 2 pose un autre type de problème. On n'a pas d'interaction bien que l'*IVI* soit très favorable et qu'il y ait trois gènes en présence. La phrase est présentée dans l'exemple 26 ci-dessous.

Exemple 26 Phrase délicate à cause de la proposition *whereas*
l'auteur suggère par l'utilisation de la conjonction *whereas* que l'activation par *achaete-scute complex* et *daughterless*, d'une part, et l'inhibition par *extramacrochaete* et *hairy*, d'autre part, sont des phénomènes indépendants.

Sensilla development is promoted by members of the achaete-scute complex and the daughterless gene whereas it is suppressed by whereas extramacrochaete (emc) and hairy.

Nous voyons que seul la présence de la conjonction *whereas* permet de dire qu'il n'y a pas interaction. Il se trouve que *whereas* ne fait pas partie des termes spécifiques. On peut faire ce constat figure 4. Cette copie d'écran correspond au même formulaire que dans la figure 2 mais dans une zone qui était « hors champ ». Dans le cas présent, le système risque donc de générer par erreur des interactions entre les gènes en présence dans la phrase.

B. CONFRONTATION ENTRE L'ANNOTATION MANUELLE ET L'ANNOTATION AUTOMATIQUE

Une autre façon d'observer si la méthode que nous proposons fonctionne sur un résumé donné, consiste à comparer directement annotation manuelle et annotation automatique. Cette comparaison vaut aussi bien pour les identifications de gènes que pour les reconnaissances d'interactions.

On modifie alors de critère de sélection des informations présentées à l'écran de façon à voir à la fois l'annotation manuelle et une partie des annotations automatiques. Cette modification se fait au niveau des requêtes jacentes aux formulaires. Une copie d'écran est présentée figure 3.

L'instantané présenté correspond à la phrase 8 du résumé précédemment étudié. On découvre une partie du formulaire qui était hors champ sur la figure 2. La phrase est donnée en haut de l'écran. On constate qu'il y a eu une erreur de segmentation de la phrase. L'annotateur a coché la case correspondante en haut à droite.

La partie centrale en vert montre les reconnaissances de définitions de gènes. Plusieurs nouveaux champs ont été rajoutés par rapport à la figure précédente. Le premier champ correspond au processus de reconnaissance qui a été mis en œuvre. Il s'agit ici une fois sur deux d'un processus automatique. La première reconnaissance n'a pas été retenue par le système d'identification automatique des gènes. Cela est indiqué par le champ *Prendre* qui n'est pas coché. La raison est indiquée dans le champ *processus* qui indique que le label, *small* en l'occurrence, n'as pas été confirmé. En effet, *small* aurait dû l'être pour être interprété car il est ambigu comme cela est indiqué dans le champ *reconnaître*. Nous voyons que l'expert n'a pas non plus reconnu cette définition en *position* 41, car il n'y a pas d'enregistrement correspondant. Les autres reconnaissances automatiques sont toutes de type *prendre* et elles ont bien été confirmées par l'annotateur.

Le sous-formulaire du bas en bleu montre les reconnaissances d'interactions faites sur la même phrase soit par l'annotateur soit par le programme. Nous n'avons présenté que les annotations automatiques faites au cours du processus *2G*. L'annotation du programme et de l'expert coïncident dans cet exemple, à ceci près que l'annotation du programme est moins précise, puisqu'elle n'indique pas le sens de l'interaction.

C. AUTRES INFORMATIONS SUR LE RÉSUMÉ

Le même formulaire de consultation des annotations donne aussi accès à d'autres informations comme présenté dans la figure 4.

Le résumé en entier est présenté en bas à droite. Juste au-dessus, on trouve des informations propres au résumé dans son entier comme le numéro *Medline*, l'*IVI* du résumé ou l'annotateur du résumé.

Le sous-formulaire, au haut à droite, présente la ou les phrases de *Flybase* qui ont un lien avec le résumé. La présence d'une description d'interaction est notée dans le champ validation. Pour plus de détail on se reportera au tableau 60 qui décrit la structure de la table des phrases extraites de *Flybase*. Cette table, dont la structure est provisoire est intitulée *IPF_PF_IRM_*.

Le sous-formulaire à gauche fournit des informations sur le calcul de l'*IVI*. Les informations présentées se rapportent à une des phrases du résumé. Dans le cas présent, il s'agit de la phrase de l'exemple 26. La colonne fréquence donne le nombre de phrases de *Flybase* qui utilisent le lemme associé au terme. Si elle est faible, cela peut expliquer une valeur inattendue de la spécificité.

Tableau 69 Interactions extraites par le programme (processus *2G*)
Le seuil choisit pour l'*IVI* est –0,1. Le tableau est listé dans l'ordre alphabétique sur la première colonne puis sur la seconde colonne. Les interactions en gras ont été confirmées par l'expert.

Gène 1	Gène 2
abdominal A (abd-A)	Ultrabithorax (Ubx)
achaete (ac)	**hairy (h)**
armadillo (arm)	**shaggy (sgg)**
armadillo (arm)	**wingless (wg)**
bicoid (bcd)	**hunchback (hb)**
bicoid (bcd)	**nanos (nos)**
bicoid (bcd)	oskar (osk)
bicoid (bcd)	**runt (run)**
bicoid (bcd)	**Serendipity delta (Sry-delta)**
brahma (brm)	**absent, small, or homeotic discs 1 (ash1)**
cactus (cact)	**dorsal (dl)**
cdc2	**Cdc37**
cdc2c	**Cyclin E (CycE)**
clift (cli)	**zeste (z)**
concertina (cta)	**folded gastrulation (fog)**
copia element (copia)	**Lighten up (Lip)**
crossveinless (cv)	yellow (y)
Cyclin A (CycA)	**cdc2**
Cyclin A (CycA)	**roughex (rux)**
decapentaplegic (dpp)	**hedgehog (hh)**
decapentaplegic (dpp)	**screw (scw)**
decapentaplegic (dpp)	**thickveins (tkv)**
decapentaplegic (dpp)	**tolloid (tld)**
decapentaplegic (dpp)	**wingless (wg)**
dishevelled (dsh)	wingless (wg)
Distal-less (Dll)	**Ultrabithorax (Ubx)**
dorsal (dl)	**decapentaplegic (dpp)**
dorsal (dl)	**Dorsal switch protein 1 (Dsp1)**
dorsal (dl)	empty spiracles (ems)
dorsal (dl)	**torso (tor)**
dorsal (dl)	**zerknullt (zen)**
doublesex (dsx)	**transformer (tra) {valide}**
doublesex (dsx)	**transformer 2 (tra2)**
DP transcription factor (Dp)	**E2F transcription factor (E2f)**
dunce (dnc)	shibire (shi)
easter (ea)	snake (snk)
EGF receptor (Egfr)	**rhomboid (rho)**
engrailed (en)	**extradenticle (exd)**
engrailed (en)	**Ultrabithorax (Ubx)**
engrailed (en)	**wingless (wg)**
abdominal A (abd-A)	Ultrabithorax (Ubx)
Enhancer of Polycomb (E(Pc))	Suppressor of zeste 2 (Su(z)2)
Enhancer of split (E(spl))	**Notch (N)**
even skipped (eve)	fushi tarazu (ftz)
extra macrochaetae (emc)	**scute (sc)**
fork head (fkh)	**trithorax (trx)**
groucho (gro)	hairy (h)
gurken (grk)	EGF receptor (Egfr)
gurken (grk)	**rhomboid (rho)**
gypsy element (gypsy)	**flamenco (flam)**
Heterogeneous nuclear ribonucleoprotein at 27C (Hr…	**P-element somatic inhibitor (Psi)**
Histone H1 (His1)	High mobility group protein D (HmgD)
huckebein (hkb)	**serpent (srp)**
huckebein (hkb)	**snail (sna)**
hunchback (hb)	**Kruppel (Kr)**
hunchback (hb)	**nanos (nos)**
inflated (if)	myospheroid (mys)
Jun-related antigen (Jra)	**sevenless (sev)**
knirps (kni)	knirps-like (knrl)
knirps (kni)	**Kruppel (Kr)**
pole hole (phl)	**EGF receptor (Egfr)**
Rac1	Cdc42
Ras oncogene at 85D (Ras85D)	rolled (rl)
sans fille (snf)	Sex lethal (Sxl)
Sex lethal (Sxl)	**male-specific lethal 1 (msl-1)**
Sex lethal (Sxl)	**transformer (tra) {valide}**
shaggy (sgg)	**wingless (wg)**
sloppy paired 1 (slp1)	sloppy paired 2 (slp2)
snail (sna)	twist (twi)
Stellate (Ste)	**Suppressor of Stellate (Su(Ste))**
Suppressor of Hairless (Su(H))	**Hairless (H)**
tolloid (tld)	tolkin (tok)
torso (tor)	**Downstream of raf1 (Dsor1)**
tube (tub)	**pelle (pll)**
white (w)	**Lighten up (Lip)**
wingless (wg)	**Notch (N)**

Tableau 70 Interactions extraites par l'annotateur

Les phrases prises en compte sont les mêmes que celles de mise en œuvre du processus *2G*. Le sens n'est pas indiqué sur ce tableau. L'ordre de présentation des données est alphabétique (colonne 1 puis colonne 2). Les interactions en gras ont aussi été extraites par le programme (processus *2G*).

Gène 1	Gène 2
achaete (ac)	**hairy (h)**
armadillo (arm)	**shaggy (sgg)**
armadillo (arm)	**wingless (wg)**
bicoid (bcd)	**hunchback (hb)**
bicoid (bcd)	**nanos (nos)**
bicoid (bcd)	**runt (run)**
bicoid (bcd)	Serendipity delta (Sry-delta)
blood element (blood)	Lighten up (Lip)
brahma (brm)	**absent, small, or homeotic discs 1 (ash1)**
cactus (cact)	**dorsal (dl)**
cdc2	**Cdc37**
cdc2c	**Cyclin E (CycE)**
clift (cli)	**zeste (z)**
concertina (cta)	**folded gastrulation (fog)**
copia element (copia)	**Lighten up (Lip)**
Cyclin A (CycA)	**cdc2**
Cyclin A (CycA)	**roughex (rux)**
decapentaplegic (dpp)	**hedgehog (hh)**
decapentaplegic (dpp)	**screw (scw)**
decapentaplegic (dpp)	**thickveins (tkv)**
decapentaplegic (dpp)	**tolloid (tld)**
dishevelled (dsh)	**wingless (wg)**
Distal-less (Dll)	**Ultrabithorax (Ubx)**
dorsal (dl)	**decapentaplegic (dpp)**
dorsal (dl)	**Dorsal switch protein 1 (Dsp1)**
dorsal (dl)	**torso (tor)**
dorsal (dl)	**zerknullt (zen)**
doublesex (dsx)	**transformer (tra) {valide}**
doublesex (dsx)	**transformer 2 (tra2)**
DP transcription factor (Dp)	**E2F transcription factor (E2f)**
EGF receptor (Egfr)	**rhomboid (rho)**
engrailed (en)	**extradenticle (exd)**
engrailed (en)	**Ultrabithorax (Ubx)**
engrailed (en)	**wingless (wg)**
Enhancer of split (E(spl))	**Notch (N)**
extra macrochaetae (emc)	**scute (sc)**
extra sexcombs (esc)	even skipped (eve)
achaete (ac)	**hairy (h)**
exuperantia (exu)	transformer 2 (tra2)
fork head (fkh)	**trithorax (trx)**
gurken (grk)	**rhomboid (rho)**
gypsy element (gypsy)	**flamenco (flam)**
hairy (h)	knirps (kni)
hairy (h)	Kruppel (Kr)
Heterogeneous nuclear ribonucleoprotein at 27C (Hr…	**P-element somatic inhibitor (Psi)**
huckebein (hkb)	**serpent (srp)**
huckebein (hkb)	**snail (sna)**
hunchback (hb)	**Kruppel (Kr)**
hunchback (hb)	**nanos (nos)**
Jun-related antigen (Jra)	Ras oncogene at 85D (Ras85D)
Jun-related antigen (Jra)	**sevenless (sev)**
knirps (kni)	**Kruppel (Kr)**
male-specific lethal 2 (msl-2)	male-specific lethal 1 (msl-1)
pole hole (phl)	**EGF receptor (Egfr)**
Sex lethal (Sxl)	**male-specific lethal 1 (msl-1)**
Sex lethal (Sxl)	**transformer (tra) {valide}**
shaggy (sgg)	**wingless (wg)**
Stellate (Ste)	**Suppressor of Stellate (Su(Ste))**
Suppressor of Hairless (Su(H))	**Hairless (H)**
torso (tor)	**Downstream of raf1 (Dsor1)**
tube (tub)	**pelle (pll)**
white (w)	**Lighten up (Lip)**
wingless (wg)	**Notch (N)**

Figure 2 Formulaire d'annotations (graphiques synthétiques)

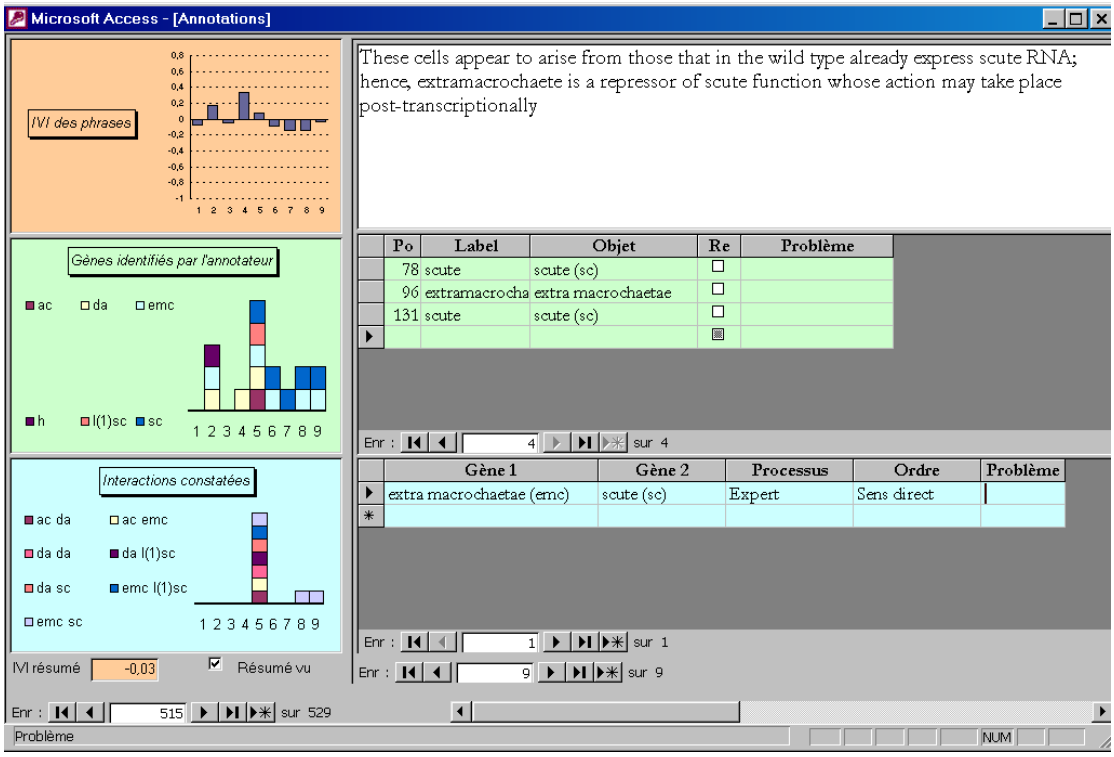


Figure 3 Formulaire d'annotations (comparaison des annotations)

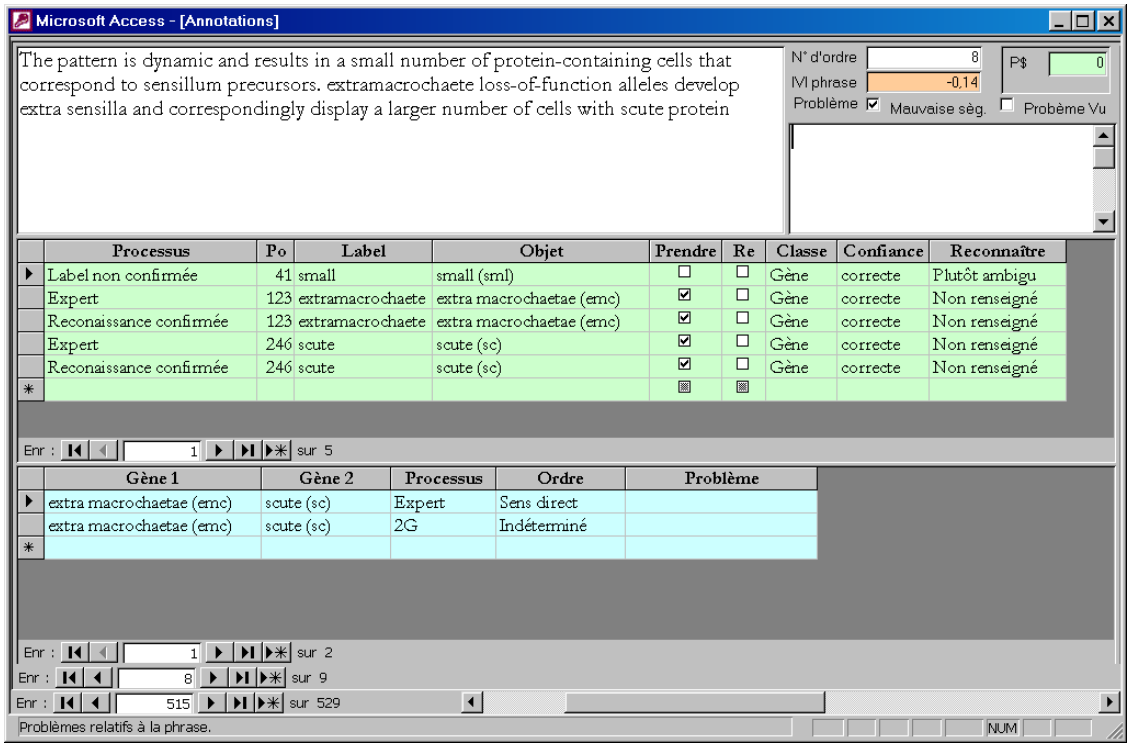


Figure 4 Formulaire d'annotation (autres informations)

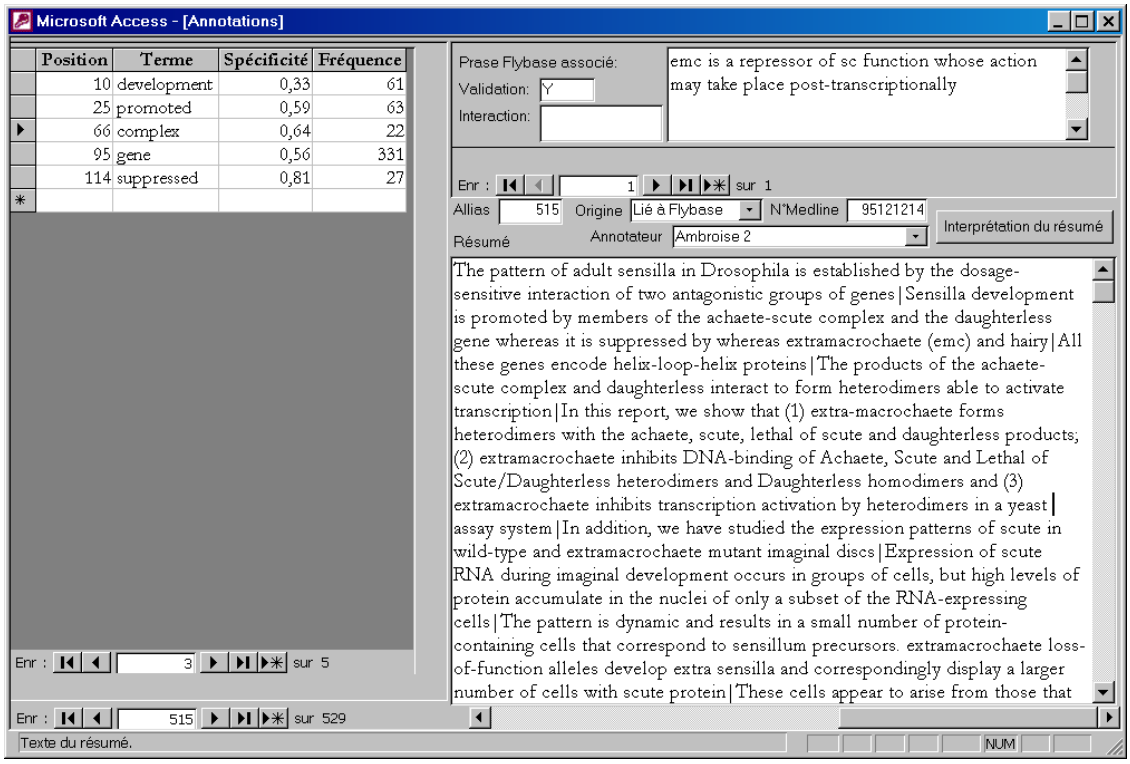
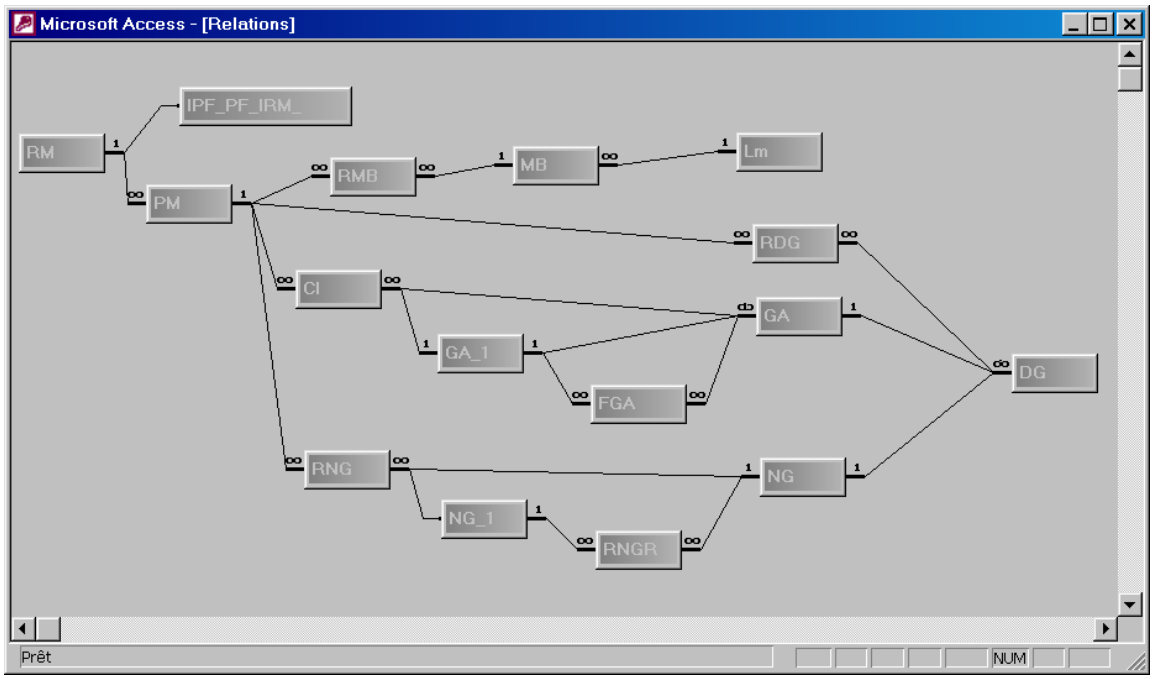


Figure 5 Schéma de la base de données

Les tables qui comptent moins de douze enregistrements ne sont pas représentées ici. Les traits représentent les relations. Pour plus de détail sur la signification de chaque relation, se reporter aux sections qui décrivent la table. Le code des tables est définit dans le tableau 38.



Chapitre 3 Évaluation et propositions d'améliorations

Nous traitons dans cette partie de l'évaluation du programme d'identification des gènes et de reconnaissance des interactions. Des propositions d'améliorations de système sont aussi incluses.

I. ÉVALUATION DU PROGRAMME D'IDENTIFICATION DES GÈNES ET NOUVELLE DIRECTIONS DE RECHERCHE

Le système d'identification des gènes a été évalué sur deux échantillons. Le premier correspond aux textes qui ont servi à mettre au point le système. Le second est constitué de textes totalement nouveaux.

A. ÉVALUATION DU SYSTÈME D'IDENTIFICATION DES GÈNES SUR L'ÉCHANTILLON A

Le système a été évalué sur l'*échantillon A*. Sur cet échantillon, 1274 identifications étaient à faire. Trois identifications n'ont pas été faites par le système et une identification faite par le système l'a été à tort. Le contexte des deux premières identifications manquées est donné dans l'exemple 27. La troisième correspond à l'interprétation du label *ras* présenté dans l'exemple 10.

Exemple 27 Reconnaissance manquée
Dans la première phrase c'est l'absence d'espace entre le label *Bic-D* (souligné) et le terme *null* qui est à l'origine de l'absence d'identification. Dans la deuxième phrase c'est la reconnaissance du terme *Musca PRI* (souligné) qui bloque celle du terme *PRI*. Le gène *Arrestin B (Arr2)*, associé à *PRI*, ne peut donc pas être reconnu. Ces deux phrases ne sont pas dans le même résumé.

A new class of Bic-Dnull alleles reveals a novel requirement for Bic-D for zygotic viability. Immunocytochemistry at the EM level revealed a distribution of both Drosophila and Musca PRI epitopes in membranous vesicular structures in the cytosol as well as in the rhabdomeric microvillar membranes where the visual pigment, rhodopsin, exists.

Le contexte de l'unique reconnaissance à tort est donné dans l'exemple 28.

Exemple 28 Reconnaissance à tort
Le terme (souligné) *Rel* a été interprété par le système comme une citation du *gène Relish (Rel).*

The Drosophila protein Dorsal (which, like the human protein NF-kappa B3, is a member of the Rel family of transcriptional activators) activates the twist gene and represses the zen gene in the ventral region of early embryos.

Ce décompte ne tient compte que des reconnaissances des seuls gènes. Les performances sur l'identification des autres types d'entité biologique n'ont pas été évaluées. En effet, seules les reconnaissances des gènes ont été contrôlées par l'annotateur.

On voit que le système est très fiable sur les résumés qui ont servi à la construction des connaissances sur la terminologie. En effet si nous les calculions sur cet échantillon, le rappel et la précision seraient supérieurs à 99 %. Nous calculerons les performances réelles du système sur un échantillon qui n'a pas servi à la mise au point du système. Nous retenons simplement que l'algorithme que nous proposons prend en compte suffisamment de critères pour arriver à une identification sans erreur ou presque. Ainsi à quelques très rares exceptions près, si le système fait une erreur, c'est que les données du dictionnaire

sont inexactes ou simplement mal-adaptées. Autrement dit, le système ne se trompe pas par manque d'intelligence mais par manque de connaissances. Le système est suffisamment subtil, mais pas suffisamment savant.

Par ailleurs, nous avons constaté sur l'*échantillon A* que le système d'identification des gènes ne fait jamais d'erreur dans le champ *Redondant*.

B. ÉVALUATION DU SYSTÈME D'IDENTIFICATION DES GÈNES SUR L'ÉCHANTILLON B ET PROPOSITIONS D'AMÉLIORATIONS

1. Performance du système d'identification des gènes sur l'échantillon B

Les données de l'échantillon A ne sont pas significatives pour évaluer le système, car ce sont elles qui ont servi à la correction des données terminologiques. Nous avons donc constitué un deuxième échantillon, l'***échantillon B***. Il a été annoté par le programme en aveugle, c'est à dire avant que nous ayons pris connaissance des textes qui s'y trouvent et fait les adaptations nécessaires dans le dictionnaire des gènes. L'échantillon B est constitué de 50 résumés. L'annotateur a effectué 408 identifications de gènes. Le programme a effectué 396 identifications de gènes. 349 identifications sont identiques. Nous pouvons donc dresser le tableau suivant :

Tableau 71 Performance du programme d'identification des gènes
L'évaluation a été faite sur un ensemble de résumés qui n'a pas été utilisé pour la mise au point du système.

Indicateur	Calcul	Valeur
Rappel	349/408	86 %
Précision	349/395	88 %

2. Un exemple de résumé annoté par le programme d'identification des gènes

Ci-dessous, nous présentons un résumé de difficulté moyenne.

Exemple 29 Résumé de difficulté moyenne pour l'identification des gènes
Les termes sur-lignés correspondent aux labels que l'annotateur a interprétés. Les termes sous-lignés correspondent aux labels interprétés par le programme. Quand le soulignement est ondulé, cela signifie que le programme a diagnostiqué qu'il ne s'agissait pas d'un gène.

We report the characterization of loss-of-function alleles of the homoeotic mutation Regulator of postbithorax (Rg-pbx) in Drosophila melanogaster. Rg- pbx is a dominant gain-of-function mutation which shows a transformation of posterior haltere to wing in the adult cuticle. This mutant phenotype mimics that of the bithorax complex lesion postbithorax (pbx). Loss-of- function alleles described here are lethal in the embryonic stage and affect the pattern of segmentation of the embryo. Examination of the terminal phenotype of null and hypomorphic alleles of Rg-pbx has shown that inactivation of the Rg-pbx gene leads to loss of the thoracic segments and the adjacent labial segment of the Drosophila embryo. An effect of the mutations is also seen in the seventh and eighth abdominal segments of embryos. The loss-of-function phenotype is similar to that described for the segmentation mutant hunchback (hb). Complementation tests show that Rg- pbx and hb are allelic. Temperature shift experiments using a temperature-sensitive loss-of-function allele show that the Rg-pbx gene product is required early in embryogenesis. We further report that the dominant Rg-pbx phenotype is sensitive to the gene dosage of another segmentation- controlling gene, fushi tarazu (ftz). Flies carrying a mutant copy of the ftz gene in trans to Rg-pbx show a dramatic enhancement of the penetrance of the homoeotic mutant phenotype. We were also able to demonstrate a suppression of the Rg-pbx phenotype by the addition of a duplication for the ftz+ gene to an Rg-pbx stock. Examination of the phenotype of ftz Rg- pbx-double-mutant embryos did not reveal a clear pattern of epistasis between the genes nor was absolute additivity of phenotype seen. A possible formal relationship between Rg-pbx, ftz, and the postbithorax (pbx) locus is proposed.

Ce résumé fait partie de l'échantillon B, ce qui signifie qu'il n'a pas été utilisé lors de la mise au point du système. Le tableau 74 permet de savoir comment le programme a interprété le résumé. Dans le même tableau sont placés à la fois les annotations du programme et de l'expert. En principe l'annotation de l'expert est immédiatement suivie d'une annotation identique du programme. C'est le cas pour la phrase 1. En revanche, au début de la phrase 2 il y a désaccord. L'annotateur a interprété le label *Rg- pbx* alors que le programme a interprété les labels *Rg* et *pbx* indépendamment. C'est l'espace après le tiret qui a induit le programme en erreur. L'erreur est facilement corrigeable en rajoutant la définition *Rg- pbx* au dictionnaire des gènes.

On remarque que les labels *labial*, *abdominal*, *similar, early et double* ont étés correctement interprétés puisque le programme a rejeté les définitions associées. En effet, la colonne *Prendre* prend la valeur *nom* dans les lignes correspondantes. Ce bon comportement est dû au fait que le problème s'est déjà posé lors de l'annotation de l'échantillon A. Ainsi nous avons déjà classé ces labels dans la catégorie des labels ambigus. Leur reconnaissance doit être confirmée pour être accepté par le programme.

L'annotateur a effectué 29 identifications de gènes sur ce résumé. Le programme en a effectué exactement le même nombre. 26 annotations coïncident. Le rappel et la précision sont donc égaux et valent tous deux 90%. Ce sont approximativement les mêmes valeurs que celles qui ont été trouvées pour l'échantillon B tout entier. C'est en ce sens, que nous pouvons dire que l'exemple ci-dessus est représentatif.

3. Inventaire des cas d'erreurs sur l'échantillon B et propositions pour les éviter

Nous avons étudié les cas d'erreurs commises par le programme sur l'*échantillon B*. Après avoir fait travaillé le programme sur l'échantillon nous avons repris chaque erreur commise afin d'en identifier les raisons. Nous avons identifié 5 catégories d'erreurs, que nous avons subdivisées en sous-catégories. Le résultat de ce travail est présenté dans le tableau 75.

Deux types d'erreurs peuvent être commises. Il s'agit soit d'une annotation faite par le programme à tort, soit d'une annotation omise par le programme. Les annotations commises à tort font baisser le taux de précision. Les annotations omises font baisser le taux de rappel. Il est important de remarquer qu'un même problème peut donc être pénalisant à la fois pour le rappel et pour la précision. Par exemple, l'introduction d'un espace après de tiret dans *Rg-pbx* va avoir deux conséquences. D'une part, l'identification de *hunchback (hb)*, qui est le gène associé à *Rg-pbx*, va être omise. D'autre part, *Ultrabithorax (Ubx)*, dont *pbx* est un synonyme, sera reconnu à tort.

Dans le tableau, nous présentons les deux types d'erreurs. Le tableau est organisé en catégories et sous catégories. L'effectif de chaque catégorie et sous catégorie est calculé. Un éventuel commentaire est indiqué avant le décompte détaillé. Le décompte détaillé concerne le label, tel qu'il est écrit dans le texte.

Nous donnons ci-dessous une analyse des catégories rencontrées, des plus fréquentes aux moins fréquentes.

La catégorie la plus fréquente, à savoir, *Variation orthographique* correspond aux variations sur des définitions de gènes déjà présentes dans *Flybase* mais qui n'ont pas été anticipées par le programme.

Les espaces après les tirets sont un problème spécifique à la source de données que nous avons utilisée. L'éditeur des cédérom a effectué un traitement sur les textes probablement pour permettre au logiciel d'interrogation de fonctionner correctement. Ce traitement consiste en l'ajout d'espace après certains tirets. Ce traitement n'est pas souhaitable pour notre application. Il faudrait soit rectifier les données, soit utiliser une autre source d'information. La rectification paraît difficile car le traitement des tirets n'est pas systématique : tous les tirets ne sont pas suivis d'espace. Il faudrait savoir quelle a été la logique de ce traitement pour pouvoir le rectifier. L'utilisation des données issues de l'internet paraît plus simple. Le traitement mis en cause n'a pas été effectué sur les données présentes sur internet. Une autre solution consisterait à modifier les données du dictionnaire pour le mettre en adéquation avec les textes. Il s'agirait d'ajouter de nouvelles définitions chaque fois qu'un tiret apparaît dans un label. Cette solution est moins élégante que les deux précédentes. Une troisième solution consisterait à supprimer tous les espaces après les tirets dans les textes. Cependant cela pourrait conduire à créer de nouveaux problèmes à d'autres endroits. Nous avons choisi de n'effectuer aucun pré-traitement des textes. Nous préférons ajouter des définitions au dictionnaire pour prendre en compte toutes les irrégularités présentes dans les textes. Le suivi des opérations est correctement assuré dans le dictionnaire alors qu'il serait difficile de l'assurer dans les textes. Ainsi, dans l'évaluation du système, nous pouvons facilement savoir quand une erreur est due à un de nos traitements.

Les parenthèses incorrectes correspondent aussi à un pré-traitement des textes dont nous disposons. Des parenthèses ont été remplacées par des crochets. Ce traitement n'est pas systématique et correspond à une logique difficile à élucider. Ce traitement introduit des dissymétries : des crochets ferment des parenthèses, des parenthèses ferment des crochets et parfois même deux parenthèses ouvrantes ne sont fermées que par un seul crochet. Les données issues de l'internet présente les mêmes défauts aux même endroits. Actuellement le problème des parenthèses incorrectes n'est pas traité automatiquement. Les irrégularités sont simplement relevées et ajoutées au dictionnaire.

Un traitement possible consisterait à remplacer tous les crochets par des parenthèses dans les textes. Comme dans le cas précédent, cela pourrait créer de nouveaux problèmes à d'autres endroits.

La catégorie espace avant et après les exposants concerne aussi un problème de formatage. Il s'agit cette fois de données erronées dans le dictionnaire et non plus dans les textes. Les exposants font partie intégrante des noms de gènes. Par exemple *su(n^a)* est le symbole d'un gène. Dans les données issues de *Flybase*, les exposants et les indices sont indiqués. L'exposant est codé dans *Flybase* par du HTML. La notion d'exposant ou d'indice est absente des résumés issus de *Medline*. C'est pourquoi nous avons supprimé le code HTML des données issues de *Flybase*. Dans la plus part des cas, cela permet effectivement de mettre en adéquation le texte et le dictionnaire. Cependant, dans certain cas, des espaces sont présents entre les balises HTML et le texte lui-même, mais ce n'est pas systématique. Ils n'ont pas été supprimés lorsque nous avons effectué la suppression des balises HTML. C'est la cause des problèmes constatés. La solution à ce problème consiste à réimporter les données issues de *Flybase* avec cette fois une procédure permettant de supprimer ces espaces.

Le cas des coupures de mots est plutôt délicat. Il faudrait disposer d'un logiciel de reconnaissance approximative de mots pour compléter automatiquement le dictionnaire. Cette complémentation suivrait le même principe de validation qu'actuellement, à savoir, une validation par le contexte. Le système d'identification des gènes que nous proposons nécessite un dictionnaire des gènes le plus complet possible. Malheureusement, les dictionnaires ne sont jamais tout à fait complets.

Nous proposons une méthode pour compléter le dictionnaire grâce à une analyse automatique des textes. Cette méthode consiste à anticiper les variations orthographiques possibles, puis à les valider sur les textes. Cette validation utilise le phénomène de répétition du même gène sous des appellations différentes. Ainsi, quand, dans le même texte, une appellation variante côtoie une appellation répertoriée, cela valide l'appellation variante. Cette méthode permet d'anticiper les variantes orthographiques dans environ 70 % des cas. Pour aller plus loin, il faudrait utiliser des techniques de reconnaissance approximative de chaînes de caractères. Ces techniques permettraient de reconnaître des définitions variantes possibles. Cependant, nous considérons que ces travaux, bien qu'utiles, font partie d'un domaine de recherche distinct, à savoir l'acquisition de connaissances terminologiques à partir de corpus.

Les cas *divers* rassemble, entre autres, des variations orthographiques portant simultanément sur la case et sur l'équivalence entre tiret et espace. Ce sont des variations qui ne sont pas prises en charge par le système.

La catégorie *manque du dictionnaire* correspond aux définitions absentes du dictionnaire. Il n'y que trois cas au total, mais les deux premiers concernent une dizaine d'occurrences chacun. Ces deux cas sont plutôt atypiques. Il s'agit des labels *DNA ligase I* et *DNA ligase II* qui sont manifestement les *noms complets* des gènes du même nom. Il s'agit d'une erreur assez grossière de *Flybase*. Il n'y avait pas de *noms complets* pour ces gènes dans la base de données. Seuls étaient présent les *symboles* et des synonymes.

Le cas du label *alpha-methyldopa hypersensitive* est plus classique. Il s'agit d'un synonyme absent du dictionnaire. Ce terme désignait bien le gène *alpha methyl dopa-resistant (amd)* dans le texte car l'auteur a précisé entre parenthèses le *symbole* du gène. De plus, une visite sur le site de

Flybase, nous apprend que le label *l(2)amd alpha-methyldopa hypersensitive* est un synonyme du gène.

Le seul traitement que nous envisageons pour cette catégorie est de compléter manuellement le dictionnaire. Cela ne permet pas d'anticiper sur de nouveaux cas.

Nous avons déjà largement présenté le problème des labels ambigus. Cette catégorie est encore assez importante mais ce n'est plus la principale source d'erreurs comme au début de nos expérimentations. Ce progrès est dû à l'accumulation d'informations sur les labels ambigus. L'annotation de nouveaux textes permet de découvrir sans cesse de nouveaux labels ambigus. Une solution définitive au problème consisterait à utiliser des lexiques de termes courants de l'anglais pour faire la liste des labels potentiellement ambigus. Cette solution paraît acceptable si l'on utilise aussi le contexte lors de l'interprétation. Les labels considérés comme potentiellement ambigus ne seraient pas totalement négligés. Lors de l'interprétation, ils ne seraient rejetés que s'ils correspondent à des *reconnaissances isolées*. Concrètement, une nouvelle catégorie serait créée dans la table *Type de reconnaissance*. Les labels appartenant à une liste de termes potentiellement ambigus seraient classés dans cette catégorie. Ils recevraient un traitement identique au label de la catégorie *plutôt ambigus*.

Cependant, il faut bien noter que cela conduirait nécessairement le système à faire des erreurs là où il n'y en avait pas précédemment. Ainsi, des labels peu ambigus dans le contexte de la génétique de la drosophile comme *hedgeog* risquerait de ne pas être interprétés comme des gènes. Il n'est donc pas sûr que cette méthode de résolution du problème serait la bonne. La technique que nous avons adoptée, qui consiste à considérer que les labels sont univoques jusqu'à preuve du contraire, a l'avantage de permettre un rappel fort.

La catégorie *Confusion entre entités* regroupe le cas du chromosome *SD chromosome* et du complexe *decapentaplegic complex*. Ces entités ont été confondues respectivement avec le gène *Sd* et avec le gène *decapentaplegic*.

Les problèmes de ce type pourraient être évités à l'avenir en utilisant un lexique de termes de biologie. Les termes du lexique qui permettraient de lever l'ambiguïté sur un label seraient inclus dans le dictionnaire.

Le tableau ci-après fait la synthèse du tableau 75.

Tableau 72 Inventaire des cas d'erreurs sur l'échantillon B (tableau de synthèse)
La colonne + indique le nombre d'erreur par excès : le programme identifier par erreur un gène. La colonne - indique le nombre d'erreur par défaut : le programme a omis d'identifier un gène. La colonne *T* donne le total. Les colonnes en grisé donnent les proportions afférentes aux effectifs qui précèdent.

Type d'erreur	-	%	+	%	T	%
Variation orthographique	35	59	12	26	47	45
Manque du dictionnaire	20	34			20	19
Label ambigu			18	39	18	17
Confusion entre entités			13	28	13	12
Divers	4	7	3	7	7	7
Total	59	100	46	100	105	100

Les effectifs sont suffisants pour permettre de faire des calculs de pourcentage. Nous voyons que la catégorie *Variation orthographique* est la plus fréquente. Elle totalise presque la moitié des erreurs (45 %). Cette catégorie provoque des erreurs des deux types. Elle est responsable de plus de la moitié des reconnaissances manquées (59 %) et d'environ un

quart des faux positifs (26 %). L'effort principal nous paraît devoir être porté sur cette catégorie.

Cette catégorie est composée de cinq sous catégories, à savoir *Variation orthographique*, *Parenthèses incorrects*, *Espace avant et après les exposants*, *Coupure de mot* et *Divers*. Les solutions envisagées pour les trois premières sous-catégories sont faciles à mettre en œuvre. Ces sous catégories représentent 35 cas sur les 45 que compte la catégorie. Un travail sur cette catégorie devrait donc facilement améliorer les performances du système.

II. ÉVALUATION DU PROGRAMME DE RECONNAISSANCE DES INTERACTIONS ET DISCUSSION

La performance de chaque processus d'annotations automatiques est évaluée par le calcul des taux de rappel et de précision.

A. EXPLICATIONS COMMUNES À TOUS LES GRAPHIQUES

Nous donnons ici quelques indications valables sur l'ensemble des graphiques présents dans cette section. Le premier exemple de graphique est donné figure 6 (ci-après). Les données correspondantes sont données dans le tableau 87.

La première colonne donne le seuil appliqué à l'*IVI* de la phrase. Le test sur l'*IVI* consiste à rejeter la reconnaissance extraite si l'*IVI* de la phrase est inférieur au seuil fixé. La première valeur du seuil, qui est -2, correspond en réalité à une extraction d'information qui ne prend pas en compte l'*IVI*. En effet, l'*IVI* ne peut par définition être inférieur à -1. Le point correspondant sur le graphique se trouve à l'extrémité droite. Il est souvent un peu en retrait par rapport aux autre points qui se trouvent eux globalement sur une ligne. Cela traduit le fait que c'est un cas limite.

A l'opposé les dernières valeurs du seuil correspondent à une sélection drastique des données à extraire. Les points correspondants se trouvent à l'extrémité gauche sur les graphiques. Les effectifs associés sont faibles. Les quotients rappel et précisions sont donc moins significatifs pour ces points. C'est ce qui explique la dispersion plus grande des valeurs à l'extrême gauche des graphiques. Les données qui correspondent à des effectifs inférieurs à dix ont été supprimés car elles ne sont pas significatives.

La deuxième colonne donne l'effectif des données qui ont été extraites automatiquement. La colonne suivante donne l'effectif des données extraites par l'expert qui a annoté les textes. Il s'agit donc de la référence. La colonne *confirmé* donne l'effectif des données extraites automatiquement qui se trouvent aussi dans l'ensemble des données extraites par l'annotateur. Le rappel et la précision sont calculés à partir de ces trois dernières colonnes.

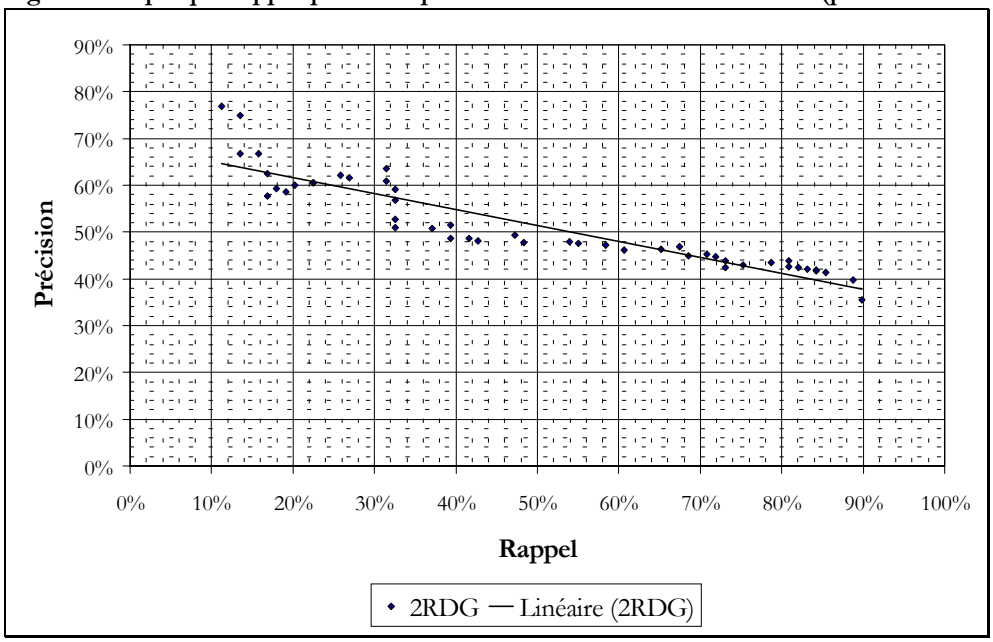
Des droites de régression ont été tracées pour faciliter la lecture des graphiques en donnant une tendance. Il ne s'agit ni de la vérification d'une hypothèse statistique, ni d'une tentative d'extrapolation.

B. STATISTIQUES SUR LES RECONNAISSANCES D'INTERACTIONS

Nous cherchons à comparer l'ensemble des reconnaissances d'interactions faites manuellement d'un part et automatiquement d'autre part.

Cette statistique n'a été faite que pour le processus de reconnaissance intitulé *2RDG* que nous décrivons en détail plus bas. Le graphe rappel-précision est donné figure 6. Les données du calcul sont fournies dans le tableau 87.

Figure 6 Graphique rappel-précision pour les occurrences d'interactions (processus *2RDG*)



L'échantillon de référence est constitué des 225 phrases qui comptent deux occurrences de gène. Les reconnaissances « cibles » sont les reconnaissances d'interactions entre gènes qui ont été faites par l'annotateur sur ces phrases.

Le processus automatique ***2RDG*** consiste à rechercher tous les couples de reconnaissance de gènes faits dans ces phrases et à inscrire l'interaction correspondante dans la table de reconnaissance des interactions.

Ce procédé d'extraction d'informations est donc exactement le même que celui qu'a utilisé PILLET. Cependant la méthode d'évaluation est différente, puisque PILLET compte des phrases pour savoir si elles contiennent ou non des interactions, alors que nous comptons des reconnaissances d'interactions pour savoir si elles sont confirmées ou non par l'annotateur.

Ce procédé de comptage ne prend pas en compte le fait qu'une même interaction puisse être manquée dans une phrase mais réussie dans une autre. Or nous comptons beaucoup sur la redondance de l'information pour obtenir des résultats satisfaisants. Nous avons donc besoin d'une évaluation des performances qui tienne compte de ce phénomène de redondance. Nous allons donc nous intéresser dorénavant non pas aux occurrences d'interaction, mais aux interactions elles-même.

C. STATISTIQUES SUR LES INTERACTIONS

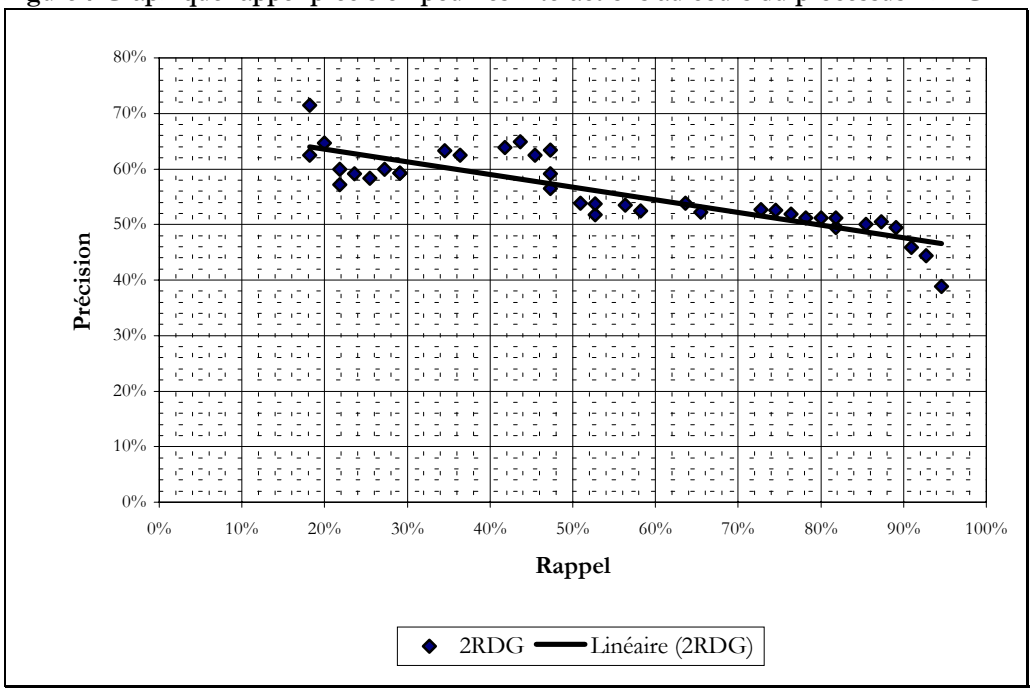
Dans cette section, nous comparons l'ensemble des interactions issues d'un processus manuel d'une part et automatique d'autre part. On ne considère plus l'extraction d'informations faite phrase par phrase mais sur l'ensemble de toutes les phrases. Dans la section B, nous comptions les reconnaissances, dans cette section nous comptons les interactions.

1. Méthodes basées sur le nombre d'occurrence de gènes dans une même phrase

a. Reconnaissance des interactions à partir des phrases qui comptent deux occurrences de gènes

Le processus d'annotation automatique concerné est intitulé *2RDG*. Il a été décrit dans la section B. Les données associées se trouvent tableau 88. La dernière colonne sera utilisée plus loin pour la comparaison de *2RDG* avec *nRDG*. Le nombre de phrases concernées par cette statistique est de 225. Le graphique correspondant se trouve figure 7.

Figure 7 Graphique rappel-précision pour les interactions au cours du processus *2RDG*



Nous remarquons que le rappel n'est au mieux que de 95%. Ce taux est atteint dans le cas de l'extraction d'informations avant prise en compte de l'*IVI*. Comment expliquer ces manques ? Dans un premier cas de figure, la phrase compte deux gènes distincts, mais elle décrit une interaction entre un gène et lui-même. Dans un deuxième cas de figure, la phrase décrit une interaction mais ne cite pas un des deux partenaires.

Nous remarquons que la précision n'est au mieux que de 65% environ. Dans ce cas, le rappel est extrêmement faible, de l'ordre de 20%. Cela signifie, que nous ne pouvons pas espérer une précision absolue (de 100%), même en tenant compte du phénomène de redondance de l'information que nous pouvons espérer sur des gros corpus.

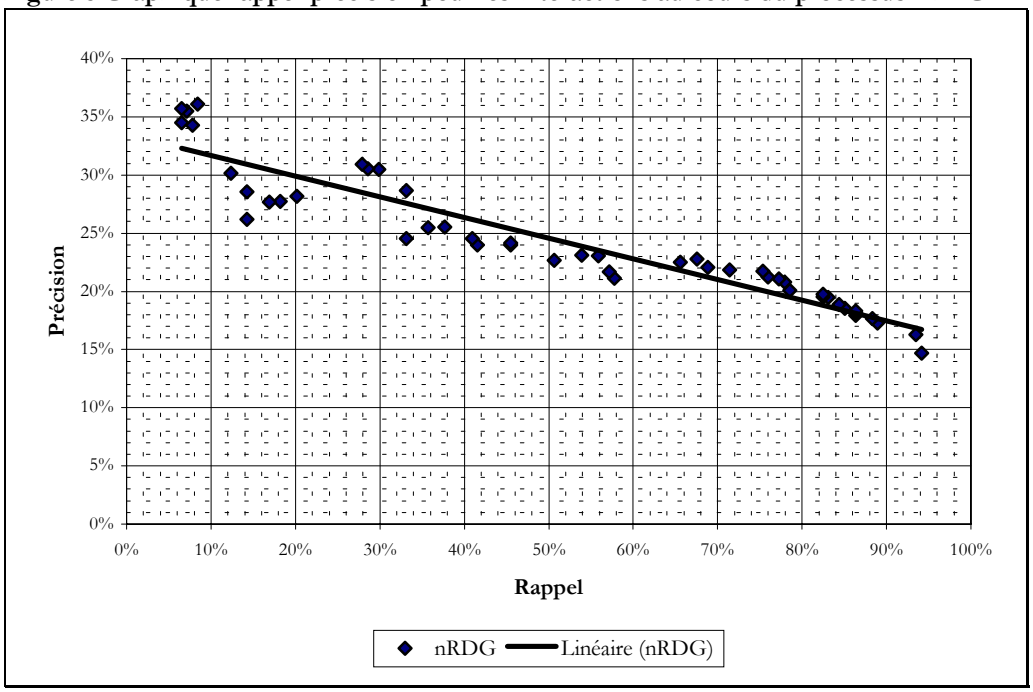
Globalement, nous remarquons que la courbe est placée assez haut sur le graphique, ce qui est positif, car cela signifie que la précision est élevée. En revanche la pente de la droite est faible, puisque bien que l'on commence assez haut, on n'atteint pas le maximum qui est 100%, mais seulement 65%. Ainsi, on voit que le principal facteur qui fait fonctionner le système d'extraction d'informations est la présence simultanée d'occurrence de gènes, la prise en compte de l'*IVI* n'apportant qu'une faible augmentation de la précision, au prix d'une forte diminution du rappel.

b. Reconnaissance des interactions à partir des phrases qui comptent plusieurs occurrences de gène

Le processus d'extraction d'information *2RDG* a l'inconvénient majeur de ne prendre en compte que les phrases qui comptent deux occurrences de gène seulement. Or on sait que de nombreuses interactions se trouvent dans les phrases qui comptent davantage d'occurrences de gène. Nous avons donné des chiffres à ce propos dans la section Chapitre 1II.B.2. Nous avons donc créé le processus de reconnaissance des interactions *nRDG*.

Le principe de la reconnaissance est le suivant : pour chaque phrase qui compte plusieurs reconnaissances de gène, pour chaque couple de reconnaissance de gène présent dans cette phrase, on crée l'interaction correspondante dans la table de reconnaissance des interactions. Le graphe correspondant se trouve figure 8, et les données correspondantes dans le tableau 89. Le nombre de phrases concernées par cette statistique est de 486.

Figure 8 Graphique rappel-précision pour les interactions au cours du processus *nRDG*



L'inconvénient de la méthode *nRDG* apparaît tout de suite : beaucoup trop d'interactions sont générées automatiquement, relativement au nombre d'interaction qui sont réellement décrites dans les phrases. On voit par exemple qu'avant intervention de l'*IVI*, près de 1000 interactions sont générées, alors que l'expert n'en a trouvé que 154. La précision ne peut, dans ces conditions, qu'être très faible.

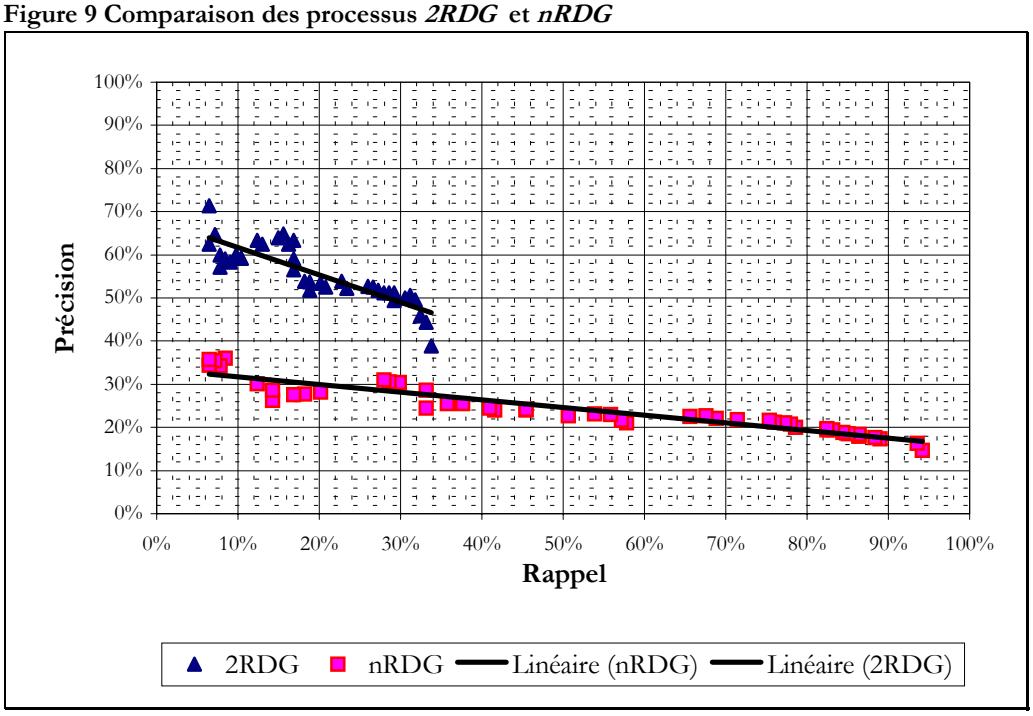
c. Comparaison des performances des méthodes basées sur le nombre d'occurrences de gènes

Nous avons vu deux processus différents : *2RDG* et *nRDG*. Chaque processus a ses avantages et ses inconvénients. Le premier est plus précis que le second, mais il ramène nécessairement moins d'interactions puisqu'il s'applique à beaucoup moins de phrases.

Plus précisément, il est clair que *nRDG* forme beaucoup de faux positifs, mais on peut espérer compenser cela en faisant une forte sélection sur les interactions en étant très

exigeant sur les *IVI* des phrases qui leur ont donné naissance. Nous allons voir précisément ce qu'il en est dans la section qui suit.. Il s'agit donc de comparer précisément *2RDG* et *nRDG* en les mettant tous les deux sur un même graphique. Pour cela nous allons considérer que *2RDG* s'applique aux même phrases que *nRDG* mais qu'il néglige de prendre en compte toutes les phrases qui comptent plus de deux occurrences de gènes. Dans ces conditions, pour le calcul des performances, les annotations de références ne sont plus les mêmes. Il faut rajouter toutes les annotations de l'expert qui ont été faites dans les phrases qui comptent plus de deux occurrences de gènes. En revanche, cela ne modifie en rien l'ensemble des annotations communes à l'expert et à la machine. Ainsi, pour un seuil donné la précision de *2RDG* reste identique, alors que le rappel change mécaniquement : il est divisé par une constante. Le nouveau rappel est indiqué dans la dernière colonne du tableau 88. Le graphique est présenté figure 9.

Figure 9 Comparaison des processus *2RDG* et *nRDG*



Nous constatons que l'effet de l'*IVI* n'est pas suffisant pour rattraper l'imprécision constitutive du processus *nRDG*. Le processus *2RDG* demeure meilleur que *nRDG* sur son domaine.

Cependant, on ne peut pas dire qu'un des processus soit globalement meilleur que l'autre. Le processus *nRDG* ne bat certes pas *2RDG* sur son domaine, qui est celui des *précision forte* et *rappel faible*. Mais il s'exerce aussi dans le domaine *précision faible* et *rappel fort* sur lequel il n'est pas concurrencé par *2RDG*. Chaque processus garde donc son intérêt.

2. Méthodes basées sur le nombre de gènes cités dans une même phrase

Nous avons vu que la présence de deux occurrences de gènes est un indice fort de la présence d'une interaction entre les gènes concernés. Cependant, on imagine bien que cet indice soit moins convaincant quand il s'agit de deux occurrences du même gène. Autrement dit, si un auteur cite conjointement deux gènes, c'est quand même un indice fort que les deux gènes ont quelques chose à faire l'un avec l'autre et donc en particulier qu'ils interagissent, tandis que si un auteur cite plusieurs fois le même gène dans la même phrase,

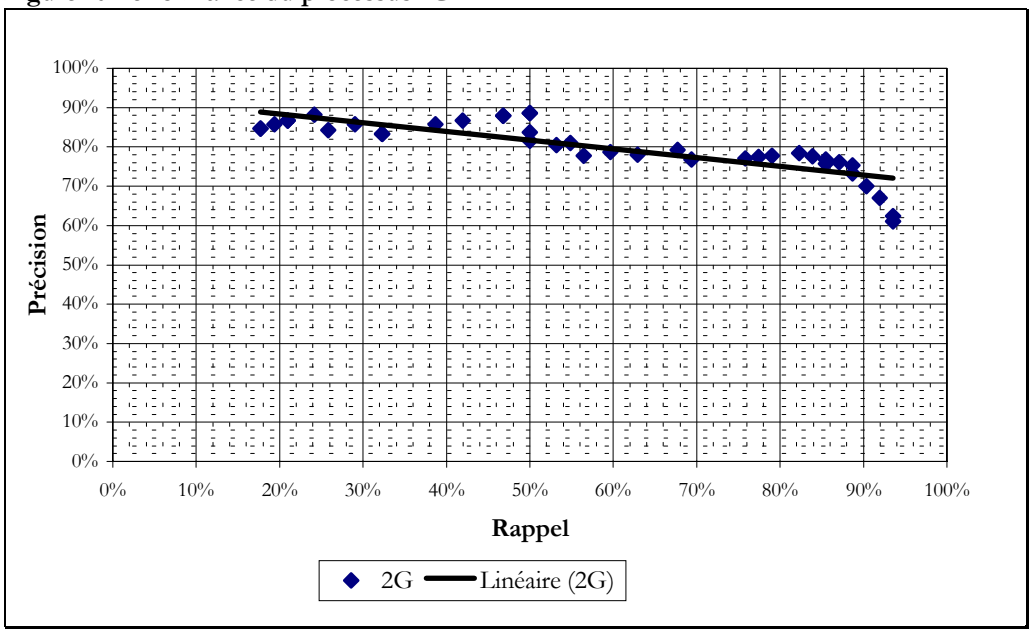
ce n'est peut-être que parce que ce gène l'intéresse. Ainsi, la méthode d'extraction d'information que nous proposons paraît être plus pertinente pour trouver des interactions entre gènes différents que pour trouver des auto-interactions. Nous allons donc reprendre nos statistiques en ne nous intéressant cette fois qu'aux interactions du premier type pour négliger les interactions du deuxième type, sur lesquels d'autres méthodes d'extraction d'interactions pourraient s'avérer plus pertinentes.

a. *Reconnaissance des interactions à partir des phrases qui citent deux gènes*

Nous nous intéressons aux phrases qui citent exactement deux gènes. Pour ces phrases nous créons tous les couples de gènes en présences. Cela nous fournit les interactions du processus ***2G***.

Ce processus prend en compte 189 phrases. Le graphique correspondant se trouve figure 10 et les valeurs se trouvent dans le tableau 90

Figure 10 Performance du processus *2G*



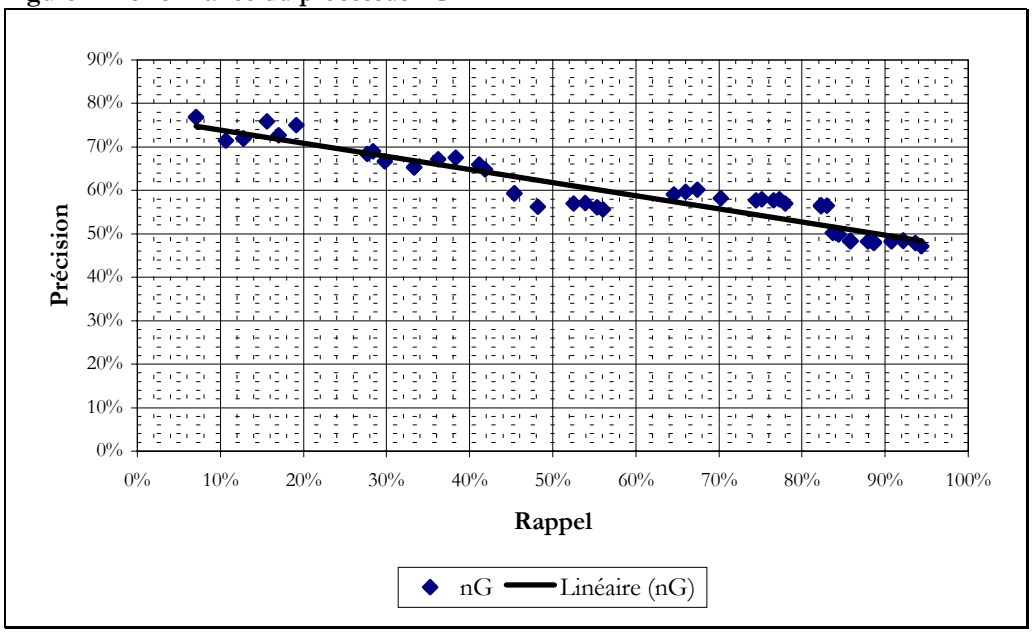
Les résultats sont effectivement meilleurs avec *2G* qu'avec *2RDG*. En effet, avant prise en compte de l'*IVI* la précision atteint 61% avec *2G* alors qu'elle n'était que de 39% pour *2RDG* avec des niveaux de rappel sensiblement égaux. Cela confirme bien l'intérêt qu'il y avait à restreindre le champ d'application de la méthode aux interactions entre gènes distincts.

b. *Reconnaissance des interactions à partir des phrases qui citent plusieurs gènes*

Exactement pour les même motifs qui nous ont conduit *nRDG* après *2RDG*, nous sommes conduits à ***nG*** après *2G*. Il s'agit de considérer les phrases qui citent plusieurs gènes, et pour chacune d'elles, de construire l'ensemble des couples de gènes qui sont cités dans cette phrase.

Ce processus prend en compte 252 phrases. Le graphique correspondant se trouve figure 11 et les données tableau 91.

Figure 11 Performance du processus *nG*

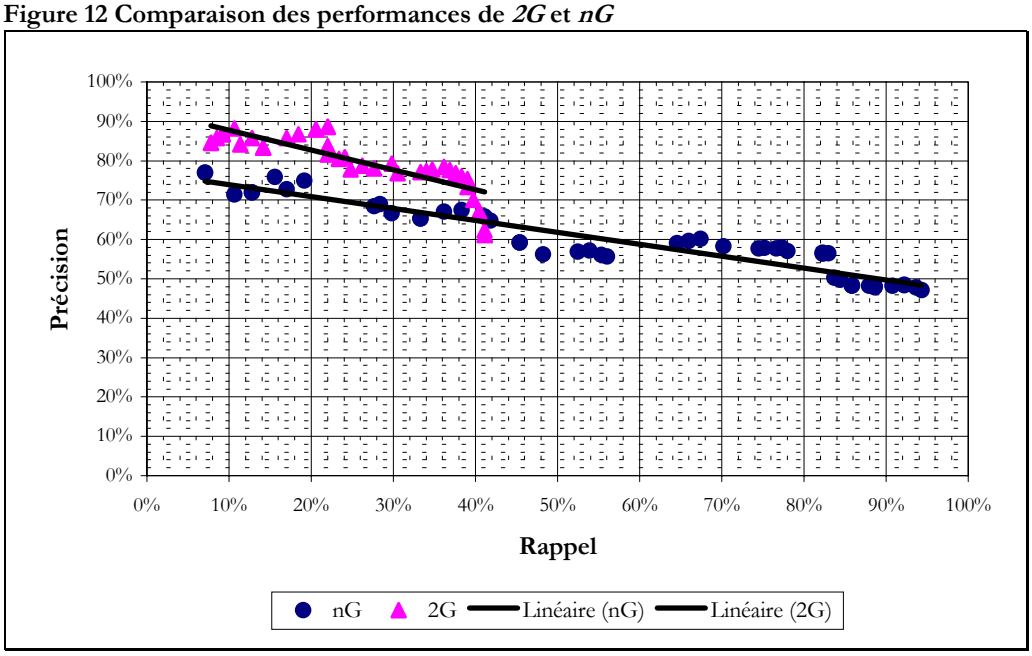


Comme lors du passage de *2RDG* à *nRDG*, on constate une diminution de la précision. Cependant elle reste relativement haute, puisqu'elle vaut au minimum 47%.

c. Comparaison des performances des méthodes basées sur le nombre de gènes cités

Pour les même raisons que précédemment avec les techniques *2RDG* et *nRDG*, il est intéressant de comparer *2G* avec *nG*. Le graphe correspondant se trouve figure 12. Les données utilisées pour tracer *nG* sont les mêmes que précédemment. Les données utilisées pour tracer *2G* figurent dans la dernière colonne du tableau 90.

Figure 12 Comparaison des performances de *2G* et *nG*



Nous constatons que là encore, le processus *2G* l'emporte là où il est concurrencé par *nG*. Cependant la différence est moindre que ce qu'elle était pour les processus *2RDG* et *nRDG*. En effet, à un niveau de rappel de 20%, le gain de précision entre *nRDG* et *2RDG* est de 25% (à lire entre les deux droites de tendance) alors qu'il est moitié moindre entre *nG* et *2G*.

3. Utilisation du nombre de fois où une interaction est reconnue automatiquement

Dans ce qui précède, nous avons considéré que la détection automatique d'une interaction pouvait se faire dans une seule phrase pour être crédible. Nous allons étudier dans cette section l'effet d'une sélection des interactions automatiques basée sur la fréquence des reconnaissances associées.

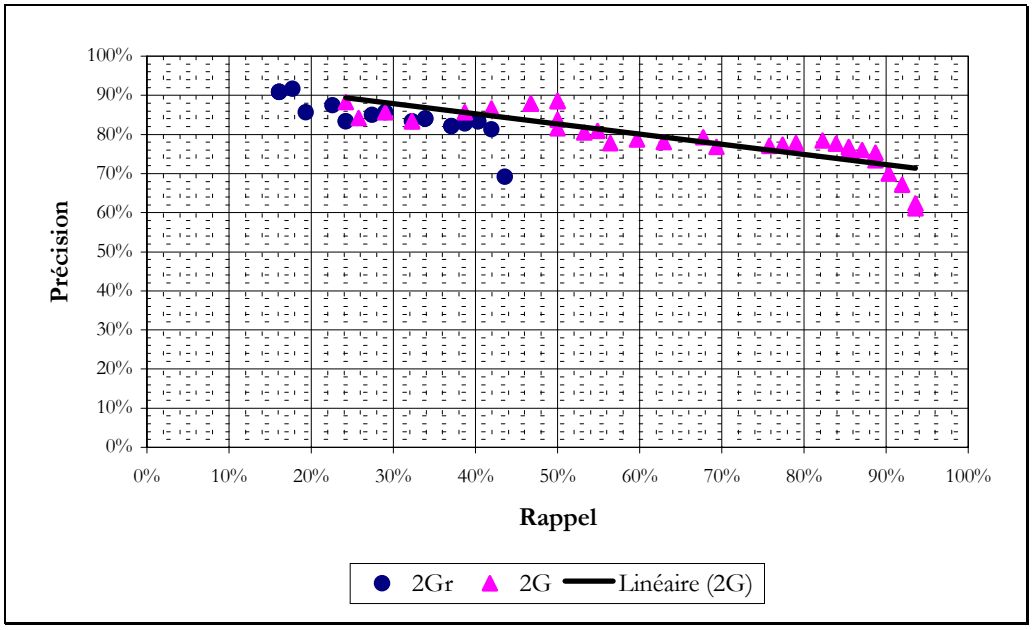
Les interactions générées automatiquement seront caractérisées de ***répétées*** si elles ont été reconnues plus d'une fois au cours du processus de reconnaissance automatique que l'on considère.

Exiger la répétition, est une façon de sélectionner les interactions, au même titre que la sélection basée sur l'*IVI*. Il s'agit de comparer les deux méthodes de sélection. Pour le faire nous proposons un graphe rappel-précision. Sur ce graphe nous proposons deux séries de données. La première correspond aux interactions extraites automatiquement sans sélection, et la deuxième série de données correspond aux seules interactions répétées. Cette comparaison a été faite pour les processus automatiques *2G* et *nG*.

a. Interactions reconnues plusieurs fois au cours du processus 2G

Le corpus de phrases qui sert à l'analyse est le même que pour le processus *2G*. Le graphe correspondant se trouve figure 13 et les données de la série ***2Gr*** se trouvent dans le tableau 92.

Figure 13 Comparaison du critère répétition avec le critère *IVI* pour le processus *2G*
La série de données *2Gr* correspond aux seules interactions répétées issues du processus *2G*.



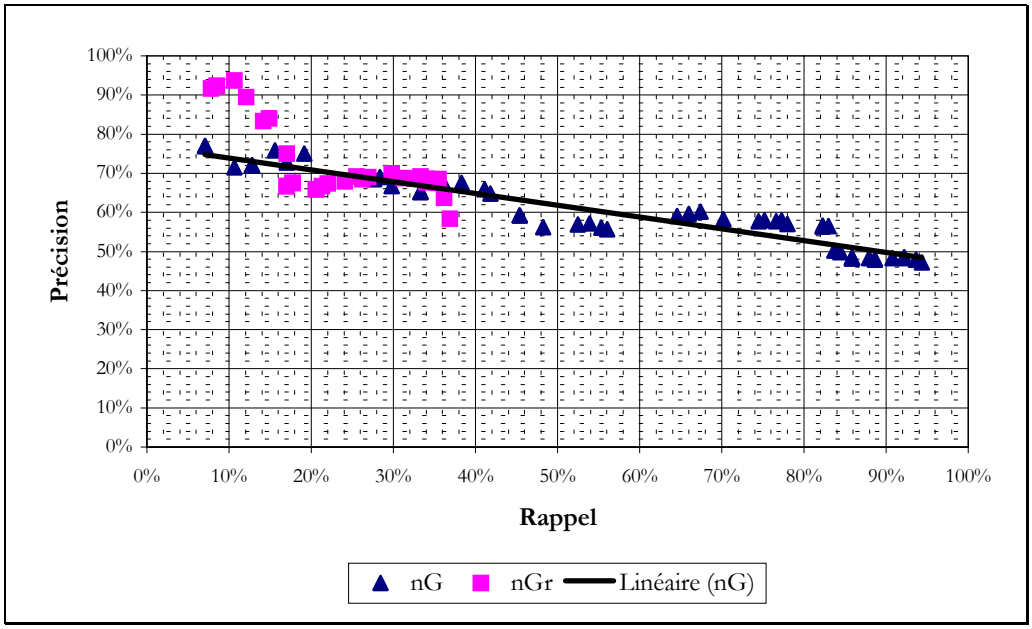
La série *2Gr* est un petit peu en dessous de la série *2G*, ce qui signifie que les performances sont légèrement inférieures. Les séries restent globalement très proches. Ainsi, on peut dire que le critère de la répétition peut rivaliser avec celui de l'*IVI*.

b. Interactions reconnues plusieurs fois au cours du processus nG

Le corpus de phrases qui sert à l'analyse est le même que pour le processus *nG*. Le graphe correspondant se trouve figure 14 et les données pour la série ***nGr*** dans le

tableau 93.

Figure 14 Comparaison des interactions extraites avec ou sans critère sur le nombre de reconnaissances associées



Les deux séries de données sont presque sur une même ligne. Cette fois-ci, la séries *nGr* n'est pas en dessous de la série *nG*. Elle parait même être au-dessus pour les points à l'extrême gauche du graphique, mais comme nous l'avons expliqué dans la section A ci-dessus, ces données sont entachées d'une grande incertitude en raison des faibles effectifs associés.

Ainsi, par rapport au cas précédent, la répétition est un critère de qualité légèrement supérieur. C'est un résultat attendu dans la mesure ou la technique d'extraction d'information *nG* donne plus de faux positifs que la technique *2G*. Il est donc naturel de penser qu'un critère basé sur l'auto-confirmation des données extraites est plus profitable dans le cas de *nG*.

c. Discussion sur la redondance de l'information sur les interactions

Nous avons vu que la sélection des interactions sur le critère de répétition donne des résultas aussi bons que l'*IVI*. Ce procédé de sélection des interactions est basé sur le fait qu'il y a redondance d'information : une interaction peut être décrite plusieurs fois dans le corpus de textes que l'on étudie. Or comme nous allons le voir ce phénomène est plutôt rare. Les résultats que nous obtenons n'en sont que plus satisfaisants.

Le tableau 73 fournit la preuve de la rareté de la redondance. En effet, on peut y voir que près des trois quarts des interactions ne sont énoncées qu'une seule fois dans les textes.

Tableau 73 Faible effectif des interactions redondantes
Les interactions ont été classées en fonction du nombre de citations –colonne redondance. Pour chaque classe on calcule l'effectif de celle-ci puis l'effectif relatif –colonne proportion. Cette statistique a été réalisée à partir de l'annotation manuelle de l'échantillon A.

Redondance	Effectif	Proportion
1	40	72 %
2	7	13 %
3	5	9 %
4	3	5 %
Total	*55*	*100 %*

D. NOUVELLES DIRECTIONS DE RECHERCHE

Les résultats que nous avons obtenus nous amènent à proposer de nouvelles directions de recherche qui pourraient être envisagé. Il s'agit d'améliorer la sélection des phrases qui semble décrire une interaction, soit par un meilleurs prise en compte du vocabulaire présent, soit en utilisant le *MeSH*.

1. Amélioration du calcul de l'IVI

Du point de vue de la biologie, les interactions n'ont pas toutes la même valeur. Certaines interactions sont plus importantes que d'autres, ne serait-ce que parce que des gènes importants y participent. Ces interactions vont être énoncées dans les textes que nous étudions un nombre important de fois. Or dans les statistiques que nous avons réalisées, toutes les interactions se valent. Il serait peut-être plus logique de mettre plus de poids aux erreurs faites sur les interactions fréquemment énoncées dans les textes. Les résultats seraient alors plus flatteurs car on a moins de chance de se tromper sur des interactions énoncées de nombreuses fois.

L'indicateur statistique que nous utilisons est extrêmement simple. La spécificité d'un terme est définie comme une proportion. Une abondante littérature existe sur les modèles mathématiques associés aux mots-clefs (SALTON *et alii*, 1983).Ces modèles servent à optimiser le calcul des poids associés aux mots clefs. Il s'agit de favoriser les mots fortement discriminants. Par exemple MARCOTTE *et alii* (2001) proposent seulement 80 mots-clefs pour discriminer des résumés. Dans leur travail, il s'agit de distinguer les résumés qui décrivent une interaction entre protéines, de ceux qui n'en décrivent pas. Le modèle mathématique qu'ils utilisent leur permet de calculer quelle est l'hypothèse la plus probable quand une série de mots-clefs est observée. Quand c'est la présence d'une interaction qui est la plus probable, le résumé est retenu pour compléter automatiquement une base de données sur les interactions entre protéines. Ce travail d'extraction d'information a été utilisé pour compléter la base ***DIP*** (voir plus loin). Les 80 mots-clefs et les poids associés à ces mots-clefs ont été calculés par apprentissage à partir de données déjà présentes dans la base de données *DIP*.

Database of Interacting Proteins (DIP) est une base de donnée multi-organismes sur les interactions entre protéines (XENARIOS *et alii*, 2000). Dans cette base de donnée relationnelle, l'information est organisée en trois tables principales. La première contient des informations sur les protéines, par exemple leurs numéros dans *SwissProt* ou *GenBank*. La deuxième table contient les informations sur les interactions proprement dites : partenaires concernés, domaine concerné, etc. La troisième contient des informations sur les conditions dans lesquels ces interactions ont été mises en évidence : références bibliographiques, type d'expérimentation mise en œuvre, etc.

Nous pourrions nous inspirer de la méthode proposée par MARCOTTE *et alii* pour calculer un *IVI* fondé sur un modèle mathématique classique. Nous pouvons espérer une amélioration des performances de cette façon, sans rien changer d'autre que la formule de calcul de la spécificité.

Les interactions génétiques ou moléculaires sont de nature variée. Chaque type d'interaction a probablement son propre vocabulaire spécifique. Nous aurions intérêt à prendre en compte ce phénomène pour déterminer le vocabulaire spécifique et pour calculer la spécificité des termes. Ainsi, nous pourrions déterminer le vocabulaire spécifique et la spécificité des termes pour chaque type d'interactions. Nous pourrions par exemple distinguer quatre types d'interactions : les interactions entre protéines, les interactions protéines--*ADN*, les interactions protéines--*ARN* et les interactions d'un gène avec ses propres produits. Nous pensons que les termes discriminants apparaîtraient mieux, car les ensembles seraient plus homogènes. De plus, cela permettrait de faire la distinction entre plusieurs type d'interactions.

2. Utilisation du MeSH pour sélectionner les résumés

Pour sélectionner les phrases qui semblent décrire une interaction, nous n'avons utilisé que les termes utilisés dans le texte des résumés. Les termes d'indexation du résumé qui se trouvent dans le champ *MeSH* pourraient être utilisés. Le ***MeSH*** permet une approche synthétique des thèmes traités dans les résumés. Un certain nombre de termes *MeSH* pourraient être choisis. Ils serviraient alors de critère de sélection des résumés. Cette méthode a été utilisée par STAPLEY et BENOIT dans leur travail sur la relation entre cooccurrences et fonction des gènes (2000).

Une autre manière d'utiliser le *MeSH* serait de faire une statistique sur les termes du *MeSH* comme nous l'avons fait sur les termes du résumé.

Nous pensons que l'utilisation du *MeSH* est une direction d'amélioration à privilégier car ce thésaurus est réputé être d'une grande qualité.

Tableau 74 Exemple d'identification de gènes dans un résumé

Le champ *confiance* est relatif à la définition désignée. Le champ *reconnaître* est relatif au label désigné. Le texte du résumé est donné Exemple 29.

N° phrase	Position	Processus	Prendre	Label	Gène	Confiance	Reconnaître
1	86	Expert	Oui	Regulator of postbithorax	hunchback (hb)	correcte	Non renseigné
1	86	Reconnaissance confirmée	Oui	Regulator of postbithorax	hunchback (hb)	correcte	Non renseigné
1	113	Expert	Oui	Rg-pbx	hunchback (hb)	correcte	Non renseigné
1	113	Reconnaissance confirmée	Oui	Rg-pbx	hunchback (hb)	correcte	Non renseigné
2	1	Expert	Oui	Rg- pbx	hunchback (hb)	correcte	Spécifié univoque
2	1	Définition ignorée	Non	Rg	rugose (rg)	non confirmée	Non renseigné
2	5	Reconnaissance confirmée	Oui	pbx	Ultrabithorax (Ubx)	correcte	Non renseigné
3	66	Expert	Oui	postbithorax	Ultrabithorax (Ubx)	correcte	Non renseigné
3	66	Reconnaissance confirmée	Oui	postbithorax	Ultrabithorax (Ubx)	correcte	Non renseigné
3	80	Expert	Oui	pbx	Ultrabithorax (Ubx)	correcte	Non renseigné
3	80	Reconnaissance confirmée	Oui	pbx	Ultrabithorax (Ubx)	correcte	Non renseigné
5	74	Expert	Oui	Rg-pbx	hunchback (hb)	correcte	Non renseigné
5	74	Reconnaissance confirmée	Oui	Rg-pbx	hunchback (hb)	correcte	Non renseigné
5	116	Expert	Oui	Rg-pbx	hunchback (hb)	correcte	Non renseigné
5	116	Reconnaissance confirmée	Oui	Rg-pbx	hunchback (hb)	correcte	Non renseigné
5	184	Label non confirmée	Non	labial	labial (lab)	correcte	Plutôt ambigu
6	67	Label non confirmée	Non	abdominal	abdominal (abd)	correcte	Plutôt ambigu
7	35	Label non confirmée	Non	similar	similar (sima)	correcte	Plutôt ambigu
7	89	Expert	Oui	hunchback	hunchback (hb)	correcte	Non renseigné
7	89	Reconnaissance confirmée	Oui	hunchback	hunchback (hb)	correcte	Non renseigné
7	100	Expert	Oui	hb	hunchback (hb)	correcte	Non renseigné
7	100	Reconnaissance confirmée	Oui	hb	hunchback (hb)	correcte	Non renseigné
8	33	Expert	Oui	Rg- pbx	hunchback (hb)	correcte	Spécifié univoque
8	33	Définition ignorée	Non	Rg	rugose (rg)	non confirmée	Non renseigné
8	37	Reconnaissance confirmée	Oui	pbx	Ultrabithorax (Ubx)	correcte	Non renseigné
8	45	Expert	Oui	hb	hunchback (hb)	correcte	Non renseigné
8	45	Reconnaissance confirmée	Oui	hb	hunchback (hb)	correcte	Non renseigné
9	100	Expert	Oui	Rg-pbx	hunchback (hb)	correcte	Non renseigné
9	100	Reconnaissance confirmée	Oui	Rg-pbx	hunchback (hb)	correcte	Non renseigné
9	132	Label non confirmée	Non	early	lodestar (lds)	à confirmer	Plutôt ambigu
9	132	Label non confirmée	Non	early	early (eay)	privilégiée	Plutôt ambigu
10	37	Expert	Oui	Rg-pbx	hunchback (hb)	correcte	Non renseigné
10	37	Reconnaissance confirmée	Oui	Rg-pbx	hunchback (hb)	correcte	Non renseigné
10	129	Expert	Oui	fushi tarazu	fushi tarazu (ftz)	correcte	Non renseigné
10	129	Reconnaissance confirmée	Oui	fushi tarazu	fushi tarazu (ftz)	correcte	Non renseigné
10	143	Expert	Oui	ftz	fushi tarazu (ftz)	correcte	Non renseigné
10	143	Reconnaissance confirmée	Oui	ftz	fushi tarazu (ftz)	correcte	Non renseigné
11	37	Expert	Oui	ftz	fushi tarazu (ftz)	correcte	Non renseigné
11	37	Reconnaissance confirmée	Oui	ftz	fushi tarazu (ftz)	correcte	Non renseigné
11	58	Expert	Oui	Rg-pbx	hunchback (hb)	correcte	Non renseigné
11	58	Reconnaissance confirmée	Oui	Rg-pbx	hunchback (hb)	correcte	Non renseigné
12	55	Expert	Oui	Rg-pbx	hunchback (hb)	correcte	Non renseigné
12	55	Reconnaissance confirmée	Oui	Rg-pbx	hunchback (hb)	correcte	Non renseigné
12	113	Expert	Oui	ftz	fushi tarazu (ftz)	correcte	Non renseigné
12	113	Reconnaissance confirmée	Oui	ftz	fushi tarazu (ftz)	correcte	Non renseigné
12	129	Expert	Oui	Rg-pbx	hunchback (hb)	correcte	Non renseigné
12	129	Reconnaissance confirmée	Oui	Rg-pbx	hunchback (hb)	correcte	Non renseigné
13	33	Expert	Oui	ftz	fushi tarazu (ftz)	correcte	Non renseigné
13	33	Reconnaissance confirmée	Oui	ftz	fushi tarazu (ftz)	correcte	Non renseigné
13	37	Expert	Oui	Rg- pbx	hunchback (hb)	correcte	Spécifié univoque
13	37	Définition ignorée	Non	Rg	rugose (rg)	non confirmée	Non renseigné
13	41	Reconnaissance confirmée	Oui	pbx	Ultrabithorax (Ubx)	correcte	Non renseigné
13	46	Label non confirmée	Non	double	double	correcte	Plutôt ambigu
14	40	Expert	Oui	Rg-pbx	hunchback (hb)	correcte	Non renseigné
14	40	Reconnaissance confirmée	Oui	Rg-pbx	hunchback (hb)	correcte	Non renseigné
14	48	Expert	Oui	ftz	fushi tarazu (ftz)	correcte	Non renseigné
14	48	Reconnaissance confirmée	Oui	ftz	fushi tarazu (ftz)	correcte	Non renseigné
14	61	Expert	Oui	postbithorax	Ultrabithorax (Ubx)	correcte	Non renseigné
14	61	Reconnaissance confirmée	Oui	postbithorax	Ultrabithorax (Ubx)	correcte	Non renseigné
14	75	Expert	Oui	pbx	Ultrabithorax (Ubx)	correcte	Non renseigné
14	75	Reconnaissance confirmée	Oui	pbx	Ultrabithorax (Ubx)	correcte	Non renseigné

Tableau 75 Inventaire des cas d'erreurs sur l'échantillon B

La colonne + indique le nombre d'erreur par excès : le programme identifier par erreur un gène. La colonne - indique le nombre d'erreur par défaut : le programme a manqué d'identifier un gène. La colonne *Tot.* donne le total.

Problèmes	-	+	Tot.
Variation orthographique	**35**	**12**	**47**
Espace après un tiret	***6***	***5***	***11***
Rg- pbx dans le texte, *Rg-pbx* dans le dictionnaire			
pbx		3	3
Rg- pbx	3		3
suppressor- of-white-apricot dans le texte, *suppressor-of-white-apricot* dans le dictionnaire			
suppressor -of-white-apricot	1		1
suppressor- of-forked dans le texte, *suppressor-of-forked* dans le dictionnaire			
suppressor- of-forked	2		2
forked		2	2
Parenthèses incorrects	***8***	***4***	***12***
su(wa] dans le texte, *su(wa)* dans le dictionnaire			
su(wa]	2		2
su[f) dans le texte, *su(f)* dans le dictionnaire			
su[f)	2		2
f		2	2
suppressor- of-forked (su(f] dans le texte (pas de parenthèse fermante)			
su(f]	2		2
f		2	2
E(wa] dans le texte, *E(wa)* dans le dictionnaire			
E(wa]	1		1
Défaut de confirmation de *Enhancer-of-white-apricot* à cause de la non reconnaissance de *E(wa]*			
Enhancer-of-white-apricot	1		1
Espace avant et après les exposants exposants	***12***		***12***
su(wa) dans le texte, *su(w a)* dans le dictionnaire			
su(wa)	9		9
Non confirmation de *suppressor-of-white-apricot* à cause de la mauvaise reconnaissance de *su(wa)*			
suppressor-of-white-apricot	2		2
E(wa) dans le texte, *E(w a)* dans le dictionnaire			
E(wa)	1		1
Coupure de mot	***5***		***5***
mastermind dans le texte, *master mind* dans le dictionnaire			
master mind	1		1
night-blind-A dans le texte, *nightblind A* dans le dictionnaire			
might-bling-A	1		1
beta 2- tubulin dans le texte, *beta2-tubulin* dans le dictionnaire			
beta 2- tubulin	1		1
hsDfd inclus le gène *Deformed (Dfd)*			
hsDfd	1		1
beta-galactosidase variente prévus de *beta galactosidase* mais confirmée dans aucun résumé			
beta-galactosidase	1		1
Divers	***4***	***3***	***7***
dipeptidase-A dans le texte, *Dipeptase A* dans Flybase			
dipeptidase-A	1		1
dipeptidase-B dans le texte, *Dipeptase B* dans le Flybase			
dipeptidase-B	1		1
87E actin gene dans le texte, *Actin 87E* dans le dictionnaire			
87E actin gene	1		1
ring et Ring interprété par erreur à cause d'une fausse confirmation			
ring		2	2
Ring		1	1
Faute de frappe dans le texte			
Sex comb on midleg dans le texte, *Sex combs on midleg* dans le dictionnaire			

Problèmes	-	+	Tot.
Sex comb on midleg	1		1
Manque du dictionnaire	**20**		**20**
Manque manifeste du dictionnaire	***19***		***19***
Les gènes *DNA ligase I (DNA-ligI)* et *DNA ligase II (DNA-ligII)* n'avait pas de nom complet dans le dictionnaire			
DNA ligase I	10		10
DNA ligase II	9		9
Synonyme absent	***1***		***1***
alpha methyl dopa-resistant dans le dictionnaire *l(2)amd alpha-methyldopa hypersensitive* sur Flybase *alpha-methyldopa hypersensitive* dans le texte			
alpha-methyldopa hypersensitive	1		1
Label ambigu		**18**	**18**
split		1	1
unpigmented		2	2
gel		2	2
se		1	1
Res		1	1
run		1	1
faint		1	1
arm		1	1
stranded		3	3
shut off		1	1
fat		2	2
Confusion avec le code d'une substance dans *EC 1.2.1.3*			
2.1		1	1
dorsal cuticle absent des termes spécifiques			
dorsal		1	1
Confusion entre entités		**13**	**13**
Confusion entre le gène *Sd* et le *SD chromosome*			
SD		10	10
Confusion entre le gènes *decapentaplegic* et le complexe *decapentaplegic* Le complexe était absent du dictionnaire			
decapentaplegic		2	2
Reconnaissance à tord de *antennapedia* dans l'expression *antennapedia and bithorax complexes*			
antennapedia		1	1
Divers	**4**	**3**	**7**
B2t elliminé sur liste par erreur			
B2t	3		3
Ordre des mots inattendu *Responder (Rsp) locus* dans le texte, *Responder locus (Rsp)* dans le dictionnaire			
Responder	1		1
Label imprécis		***3***	***3***
reconnaissance à tord d'*alphaTubulin84B (alphaTub84B)* par *alpha-tubulin*			
alpha-tubulin		3	3
Total	***59***	***46***	***105***

Partie 3

Conclusion

Chapitre 1 Bilan du travail

Les connaissances sur la génétique sont recensées, organisées et structurées dans des encyclopédies électroniques ou dans des banques de résultats d'expériences. Ces informations, pour être exploitées au mieux, doivent être reliées aux textes des publications qui leur correspondent. Il s'agit, soit de reconnaître dans des textes des objets décrits dans des encyclopédies électroniques, soit de rechercher des publications apportant des informations sur des résultats d'expériences.

Notre travail comporte deux volets. Le premier consiste à coupler l'encyclopédie électronique *Flybase* avec la base de données bibliographique *Medline*. Il s'agit d'identifier dans *Medline* des gènes décrits dans *Flybase*. Cela revient à indexer les résumés à l'aide des noms de gènes standards qui sont donnés par *Flybase*. Le deuxième volet consiste à construire une base de données sur les interactions génétiques ou moléculaires à partir d'un ensemble de résumés de publications. Il s'agit d'extraire des informations sur les interactions génétiques ou moléculaires à partir de résumés issus de *Medline*. Cette extraction d'informations peut servir à l'interprétation d'expériences sur les interactions génétiques ou moléculaires.

Pour parvenir à identifier les gènes dans les résumés, nous avons confronté les informations contenues dans l'encyclopédie *Flybase* avec les textes de *Medline*. Les insuffisances des données présentes dans *Flybase* ont pu être mises en évidence. Des corrections sur les données terminologiques ont été entreprises. En particulier, quand un nom de gène entre dans la composition d'un terme qui n'est pas un nom de gène, ce dernier a été ajouté au lexique de façon à éviter toute confusion. Des informations d'un type nouveau ont aussi été adjointes. En particulier, l'ambiguïté de certains noms de gènes a été évaluée et des priorités ont été données quand un terme renvoie à plus d'un gène. Ceci nous a amené à structurer les données issues de *Flybase* dans une base de données relationnelle. Dans cette base, la distinction est claire entre les informations relatives aux gènes et les informations relatives aux termes qui les désignent.

Ce travail nous a permis d'obtenir un ensemble de 108 résumés annotés. Cette annotation s'est faite de façon semi-automatique, mais le résultat final a été entièrement validé par un expert du domaine. A côté de cela, nous avons obtenu une base de données terminologiques qui a été, elle aussi, complétée de façon semi-automatique et qui a été validée par un expert. Cette validation garantit que les informations sont correctes au sens où elles rendent compte, parfaitement ou presque, de l'usage qui a été constaté dans l'échantillon de 108 résumés.

Nous avons mis au point un algorithme d'identification des gènes. Cet algorithme prend en compte le contexte. En particulier, quand plusieurs noms d'un même gène sont présents dans un résumé, cela constitue un indice qui est utilisé dans le cas où un terme serait ambigu. Le système (bases de données terminologiques plus algorithme) a été testé sur un deuxième échantillon, faisant apparaître des taux de rappel et de précision, respectivement de 85 et 87 %.

Le système permet l'import d'informations depuis *Flybase*. Ainsi, est-il possible d'actualiser le dictionnaire terminologique des gènes. Ceci est rendu possible par une mise en mémoire des corrections faites sur les données. Les données erronées ne sont pas supprimées mais

seulement invalidées. Ainsi, après le nouvel import, il ne sera pas nécessaire de contrôler à nouveau ces données pour les invalider une nouvelle fois.

La méthode que nous proposons permet d'améliorer les inventaires terminologiques existants. Elle permet d'étudier précisément l'utilisation qui est faite de la nomenclature. Nous avons pu, par exemple, comparer la fréquence d'utilisation des différents types de noms utilisés pour désigner les gènes, à savoir : les *symboles*, les *noms complets* et les synonymes.

Pour le deuxième volet de notre travail, nous avons utilisé un résultat déjà obtenu dans notre laboratoire par Pillet. Il s'agit d'une méthode qui permet de détecter les phrases qui décrivent une interaction. Cette méthode repose sur le calcul d'un indice de pertinence appelé *IVI*. Cet indice est calculé en repérant un certain nombre de termes dans la phrase et en calculant la moyenne des coefficients associés à ces termes. Les termes et les coefficients ont été déterminés par Pillet. Ce travail a été effectué sur un corpus de textes distinct du nôtre.

Nous avons utilisé l'identification des gènes qui a été mise en œuvre lors du premier volet. Les deux informations (présence d'un certain vocabulaire et présence de tel ou tel gène) ont été combinées de façon à extraire une liste d'interactions potentielles. Cette liste a été comparée à la liste des interactions effectivement observées dans le corpus.

Les résultats obtenus sont intéressants au regard de la simplicité du principe appliqué. Dans le cas où exactement deux gènes seraient cités, le taux de rappel atteindrait 89 % pour un taux de précision de 75 %. Dans le cas de phrase plus complexe où un nombre quelconque (supérieur à deux) de gènes distincts seraient cités, le taux de rappel atteindrait 82 % pour un taux de précision de 57 %.

En plus de la méthode, nous disposons maintenant de 108 résumés richement annotés. Pour chaque phrase, la liste des interactions décrites est consignée dans une base de données. Cet ensemble de résumés forme un corpus d'exemples intéressant pour les tâches d'extraction d'informations.

Grâce à cette annotation, nous avons obtenu des résultats statistiques très intéressants. Par exemple, nous pouvons dire qu'une interaction est rarement décrite dans plusieurs phrases : 72 % des interactions ne sont décrites que dans une seule phrase. Autre exemple, les descriptions d'interactions se trouvent préférentiellement dans les phrases qui comptent plus de deux occurrences de gènes : seulement 31% des descriptions d'interactions sont issues d'une phrase qui contient exactement deux occurrences de gènes.

Chapitre 2 Améliorations envisagées et nouvelles directions de recherche

I. TRANSFORMATION DU PROTOTYPE EN UN LOGICIEL CONVIVIAL

Notre base de données est un prototype. Elle demande à l'utilisateur un apprentissage pour comprendre comment accéder, modifier et rechercher les différentes informations. Son maniement nécessite de connaître la façon dont les données sont organisées. Il est souhaitable que, dans le futur, cette base de données soit utilisée par un public plus large que moi-même et les quelques personnes qui ont annoté les textes. Des adaptations seront nécessaires. Il s'agit de rajouter des interfaces qui guident l'utilisateur dans ses manipulations. Ces interfaces doivent prévoir des garde-fous pour empêcher des erreurs de manipulations qui peuvent aboutir à une corruption de la structure ou des données. Des fonctionnalités d'import sont aussi à prévoir pour les résumés et pour les données terminologiques de *Flybase*.

II. COUPLAGE AVEC DES RÉSULTATS D'EXPÉRIENCES

Nous proposons une méthode pour obtenir des informations sur les interactions génétiques ou moléculaires. Cependant, d'autres méthodes existent. Ces méthodes sont basées sur l'analyse de données d'expressions obtenues en masse. Nous pensons, en particulier, aux expériences sur *puces à ADN*. Nous pensons que notre système pourrait avantageusement être adapté à l'exploitation de ces résultats d'expériences. Il s'agirait de coupler une base de données de résultats d'expériences à une base de données bibliographique, en l'occurrence *Medline*. C'est une idée qui a été proposée par DICKERSON *et alii* (2001) qui travaillent sur un système d'extraction d'informations sur les réseaux métaboliques à partir de résumés de publications issus de *Medline* ou de la base de donnée bibliographique *Agricola*.

Les informations bibliographiques apporteraient vraiment "un plus" aux données expérimentales d'expression. Ces données fournissent des présomptions de relations entre les gènes mais elles ne donnent pas de preuves définitives. Cette preuve définitive est à rechercher dans une expérience *in vivo*. Ce sont précisément des expériences de ce type qui sont relatées dans la littérature. Ainsi, associer des données d'expression à des publications serait un moyen pratique de vérification.

Il nous semble que, d'une manière générale, les systèmes d'extractions d'informations à partir de textes donnent des résultats insuffisants pour envisager une production en masse de données de qualité. En revanche, si l'on se restreint à l'analyse des gènes qui ont une raison particulière d'être en interaction, les performances peuvent devenir acceptables. Ainsi, nous proposons de coupler les systèmes d'extraction d'informations à partir de textes à d'autres systèmes d'obtention d'informations sur les interactions.

Dans ce cas, les textes relatant une éventuelle interaction seraient sélectionnés par le critère de la cooccurrence et ils seraient classés par ordre de pertinence décroissante grâce à l'*IVI*.

Ce couplage entre bases de données de résultats d'expériences et bases de données bibliographiques compléterait le couplage que nous avons réalisé d'une encyclopédie électronique avec une base de données bibliographiques. Ainsi, nous aurions deux exemples

complémentaires de couplage de bases de donnée factuelles avec une base de données bibliographique.

III. UTILISATION DANS D'AUTRES DOMAINES D'APPLICATIONS

Le système que nous proposons, en raison de sa simplicité, est généralisable à d'autres domaines d'applications. Il est adapté à tout système de recherche et d'extraction d'informations sur les relations qu'entretiennent des objets techniques. De plus, les temps de traitement informatique sont réduits. Le système est donc adapté à une utilisation sur des données volumineuses et en constante évolution. Nous pensons en particulier aux données issues de l'internet, qui se prêtent mal à des traitements linguistiques sophistiqués.

LISTE DES TABLEAUX, FIGURES, EXEMPLES ET ÉQUATIONS

INDEX DES TERMES

BIBLIOGRAPHIE

ACHARD F, BARILLOT E. Virgil : a database of rich links between GDB and GenBank. *Nucleic. Acids Res.*, 1998, vol. 26, n°1, p. 100-101 22

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol, 1990, vol. 215, n°3, p. 403-410 17

ANDRADE M, VALENCIA A, Automatic extraction of keywords from scientific text : application to the knowledge domain of protein families. *Bioinformatics*. 1998, vol. 14, n°7, p. 600-607 33

ANDRADE MA, BORK P. Automated extraction of information in molecular biology. *FEBS Letters*. 2000, n° 476, p. 12-17 34

ANDRADE MA, VALENCIA A. Automatic annotation for biological sequences by extraction of keywords from MEDLINE abstracts. Development of a prototype system. *Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology (ISMB 97*), 1997, p. 25-32 33

ANDRADE MA. Tools for automated protein annotation. CASADIO R, MASOTTI L. *Protein Sequence Analysis in the Genomic Era*. Bologna (Italia) : CLUEB, 2001. [On line] <http://www.embl-heidelberg.de/~andrade/papers/boss99/chapter.html> 34

ANDRADE Miguel A, SANDER C. Bioinformatics : from genome data to biological knowledge. *Current Opinion in Biotechnology*. 1997, vol. 8, n° 6, p. 675-683 15

ATTWOOD Teresa K. Genomics : The Babel of bioinformatics. *Science*, 2000, vol. 290, n°5491, p. 471-473 16

BAIROCH Amos, APWEILER Rolf. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res*. 2000, vol. 28, n° 1, p. 45-48 18

BARRÉ R, LAVILLE F, TEIXEIRA N, ZITT M. L'Observatoire des sciences et des techniques : activités - définition – méthodologie. *Solaris*, n° 2, Presses Universitaires de Rennes, 1995. 11

BIKEL Daniel M, MILLER Scott, SCHWARTZ Richard, WEISCHE-DEL Ralph. Nymble : a high-performance learning name-finder. *Proceeding of the 5th Conference on Applied Natural Language Processing (ANLP'97)*, Washington. 1997. p. 194-201. ACL. Accessible en ligne <URL : http://xxx.lanl.gov/abs/cmp-lg/9803003> 20

BLASCHKE C, ANDRADE MA, OUZOUNIS C, VALENCIA A. Automatic extraction of biological information from Scientific text : protein-protein interactions. *Proceedings of the seventh international conference on Intelligent Systems for Molecular Biology (ISMB 99)*, 1999, n°7, p. 60-67 26, 34, 37

CATALANO D, LICCIULLI F, D'ELIA D, ATTIMONELLI M. Update of KEYnet : a gene and protein names database for biosequences functional organisation. *Nucleic Acids Research*, 2000, Vol. 28, N° 1, p. 372-373 30

CHEE M, YANG R, HUBBELL E. Accessing genetic information with high-density DNA arrays. *Science*. 1996, vol. 274, n° 5287, p. 610-614 17

COLLIER N, PARK HS, OGATA N, TATEISHI UY, NOBATA C, OHTA T, SEKIMIZU T, IMAI H, IBUSHI K, TSUJUII J. The GENIA project : corpus-based knowledge acquisition and information extraction from genome research papers, *Proceedings of the European Association for Computational Linguistics (EACL) conference*, 1999 38

DAVIDSON Duncan, APWEILER Rolf. Meeting the challenge of building gene function databases. Compte rendu de l'atelier *HUGO/EU Workshop*, Hinxton, Cambridge, UK, Mai 1999. [On Line] sans lieu : éditeur inconnu, sans date. URL : http://www.hgu.mrc.ac.uk/Research/Reports/Genefunc/report.htm 17

DICKERSON JA, BERLEANT D, COX Z, QI W, ASHLOCK D, WURTELE E. Creating metabolic network models using text mining and expert knowledge. *Atlantic Symposium on Computational Biology*. Genome Information Systems & Technology, 2001. 156

DISCALA C, BENIGNI X, BARILLOT E, VAYSSEIX G. DBcat : a catalog of 500 biological databases. *Nucleic Acids Research*. 2000, vol. 28, p. 8-9 22

DOU Henri. *Veille technologique et compétitivité*. Paris : Dunod, 1995. 234 p. 11

ENRIGHT AJ, ILIOPOULOS I, KYRPIDES NC, OUZOUNIS CA. Protein interaction maps for complete genomes based on gene fusion events. *Nature*. 1999, vol. 402, n° 6757, p. 86-90 18

FALOUTSOS C, OARD DW. *A Survey of Information Retrieval and Filtering Methods*. Technical Report CS-TR-3514, Department of Computer Science, University of Maryland, College Park, 1995. 19

FIELDS S ,SONG OK. A novel genetic method to detect protein-protein interactions. *Nature*. 1989, n° 340, p. 245-246 17

FITCH WM. Distinguishing homologous from analogous proteins. *Systematic Zoology*. 1970, vol. 19, n° 2, p. 99-113 17

FUJIBUCHI W, GOTO S, MIGIMATSU H, UCHIYAMA I, OGIWARA A, AKIYAMA Y, KANEHISA M. DBGET/LinkDB : an integrated database retrieval system. *Pacific Symp. Biocomputing*. 1998. p. 683-694 22

FUKUDA K, TSUNODA T, TAMURA A, TAKAGI T. Toward information extraction : Identifying protein names from biological papers. *Proceedings of the Pacific Symposium on Biocomputing*, 1998, p. 707-718. 28

GERSTEIN M., JANSEN R. The current excitement in bioinformatics-analysis of whole-genome expression data : How does it relate to protein structure and function ? *Current Opinion in Structural Biology*. 2000, vol. 10, n° 5, p. 574-584 17

GOUJON Bénédicte. *Utilisation de l'exploration contextuelle pour l'aide à la veille technologique : Réalisation du système informatique VIGITEXT*. Thèse science (traitement automatique du langage naturel) : Université Paris-Sorbonne, avril 2000. 550 p. 19

HIETER P, BOGUSKI M. Functional genomics : it's all you read it. *Science*. 1997, n°278, p. 601-602 15

HISHIGAKI Haretsugu, ONO Toshihide, TANIGAMI Akira, TAKAGI Toshihisa. Interaction-based Analysis of Protein Function and Network. *Procedings of Genome Informatics conference (GIW'99) n°10*, 1999, p. 257-258 38

HUMPHREYS B L, LINDBERG DAB, SCHOOLMAN HM, BARNETT GO. The Unified Medical language System : An informatics research collaboration. *Journal of the American Medical Informatics Association*, 1998, vol. 5, n° 1, p. 1-13 27

HUNTER Lawrence. Molecular Biology for Computer Scientists. In Hunter Lawrence. *Artificial Intelligence and Molecular Biology*. Cambrigde : AAAI Press, 1993, 500 p. ISBN 0-262-58115-9 (épuisé). En ligne : http://www.aaai.org//Library/Books/Hunter/hunter.html 13

JACQUEMIN Christian, ZWEIGENBAUM Pierre. *Traitement automatique des langues pour l'accès au contenu des documents*. In : Le document en sciences du traitement de l'information, Jacques LE MAITRE, Jean CHARLET et Catherine GARBAY (Réd.), chapitre quatrième, p. 71-109. Toulouse : Cépadues, septembre 2000. 20

JAKOBIAK François. *L'intelligence économique en pratique*. Paris : Les Éditions d'Organisation, 1998. 312 p. 11

JENSSEN TK, ÖBERG LMJ, ANDERSSON ML, KOMOROWSKI J. Methods for large-scale mining of networks of human genes. *Proceedings of the SIAM Conference on Data Mining (SDM 2001)*. Chicago, IL, USA, 2001 36

KANEHISA Minoru. *Post-genome Informatics*. Oxford, UK : Oxford University Press, 2000. 158 p. ISBN 0-19-850326 16

KARP PD. Database links are a foundation for interoperability. *Trends Biotechnol*. 1996, vol 14, n°8, p. 273-279 22

LICCIULLI F, CATALANO D, D'ELIA D, LORUSSO V, ATTIMONELLI M. KEYnet : a keywords database for biosequences functional organization. *Nucleic Acids Research*, 1999, Vol. 27, N° 1, p. 365-367 30

MAGRI MH, SOLARI A , RERAT K. Les périodiques scientifiques d'audience internationale au travers du Journal Citation Reports : analyse du système d'évaluation de l'ISI. Application à l'étude de la production de l'INRA. L'information scientifique et technique : nouveaux enjeux documentaires et éditoriaux. Colloque INRA, Tours, octobre 1996. Paris : INRA, 1997. p. 71-86 11

MARCOTTE Edward M, XENARIOS Ioannis, EISENBERG David. Mining literature for protein–protein interactions. *Bioinformatics*. 2001, vol. 17, n° 4, p. 359-363 148

MARTINET B. MARTI Y. *L'intelligence économique : les yeux et les oreilles de l'entreprise*. Paris : Éditions d'Organisation, 1995 11

MASYS Daniel R, WELSH John B, FINK J. Lynn , GRIBSKOV Michael , KLACANSKY Igor, CORBEIL Jacques. Use of keyword hierarchies to interpret gene expression patterns. *Bioinformatics*. 2001, vol. 17, n° 4, p. 319-326 35

MENDOZA L, THIEFFRY D, ALVAREZ-BUYLLA ER. Genetic control of flower morphogenesis in Arabidopsis thaliana : a logical analysis. *Bioinformatics*, 1999, vol. 15, p. 593-606 40

MUC-6 : *Proceedings of the Sixth Message Understanding Conference*. Columbia, Mary-land : Morgan Kaufmann, 1996 20

NAKAO Mitsuteru, BONO H, KAWASHIMA S, KAMIYA T, SATO K, GOTO S, KANEHISA M. Genome-scale Gene Expression Analysis and Pathway Reconstruction in KEGG. *Genome Informatics*, 1999, vol. 10, p. 94-103 17

NG SK, WONG M. Toward routine automatic pathway discovery from on-line scientific text abstracts. *Genome Informatics 1999*. Edited by ASAI K, MIYANO S, TAKAGI T. Tokyo : Universal Academy Press, 1999. p. 104-112 38

OHTA Y, YAMAMOTO Y, OKAZAKI T, UCHIYAMA I, TAKAGI T. Automatic Construction of Knowledge Base from Biological Papers, *Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology (ISMB'97)*, 1997, p. 218-225 29

OLIVEROS Juan Carlos, BLASCHKE Christian, HERRERO Javier, DOPAZO Joaquin, VALENCIA Alfonso. Expression Profiles and Biological Function. *Proceedings of the Eleventh Workshop on Genome Informatics (GIW2000)*, Tokyo, 2000. 34

ONO Toshihide, HISHIGAKI Haretsugu, TANIGAMI Akira, TAKAGI Toshihisa. Automated extraction of information on protein–protein interactions from the biological literature. Bioinformatics, 2001, vol. 17, p. 155-161 27, 30, 38

ONO Toshihide, HISHIGAKI Haretsugu, TANIGAMI Akira, TAKAGI Toshihisa. Automatic Extraction of Information on Protein-Protein Interaction from Scientific Literature. *Genome Informatics*, 1999, n° 10, p. 296-297 38

PILLET V, ROUDANI B, QUONIAM L, JACQ B. Extraction automatique et représentation graphique de données biologiques : les interactions génétiques et moléculaires. *Actes du colloque VSST'9*8, 1998, p. 215-226. 37

PILLET Violaine. *Méthodologie d'extraction automatique d'information à partir de la littérature scientifique en vue d'alimenter un nouveau système d'information. Application à la génétique moléculaire pour l'extraction d'information sur les interactions*. Thèse sci. : Science de l'information et de la communication, 2000, 128 p., 00AIX30002 8, 39

PROUX D, RECHENMANN F, JULLIARD L, PILLET V, JACQ B. Detecting Gene Symbols and Names in Biological Texts : A First Step toward Pertinent Information Extraction. *Proceding of Genome Informatics 1998 (GIW'98)*. MIYANO S, TAKAGI T. Tokyo, Japan : Universal Academy Press, 1998. p. 72-80. 28

PROUX Denys. *Muninn : une stratégie d'extraction d'informations dans des corpus spécialisés par application de méthodes d'analyse linguistique de surface et de représentation conceptuelle des structures sémantiques*. Thèse informatique : faculté des sciences et techniques, université de Bourgogne, 2001, 211 p. 40

QI W, BERLEANT D. PathBinder : a system for mining Medline for protein interactions. *Joint Bioinformatics Workshop*, 3-4 Nov. 2000, Iowa City. 26

REBHAN M, CHALIFA-CASPI V, PRILUSKY J, LANCET D. GeneCards : A novel functional genomics compendium with automated data mining and query reformulation support. *Bioinformatics*, 1998, vol. 14, p. 656-664 22

RINDFLESCH TC, HUNTER L, ARONSON AR. Mining molecular binding terminology from biomedical text. *Proc AMIA Symp*. 1999; p. 127-131. 29

RINDFLESCH TC, HUNTER L, ARONSON AR. Mining molecular binding terminology from biomedical text. *Proceedings of the AMIA Symp.*, 1999, p. 127-131 27

RINDFLESCH TC, TANABE L, WEINSTEIN JN, HUNTER L. EDGAR : extraction of drugs, genes and relations from the biomedical literature. *Proceedings of the Pac Symp of Biocomput*, 2000, p. 517-528 27, 37

ROCHA Eduardo PC. *Analyse exploratoire des génomes bactériens*. Thèse de génétique cellulaire et moléculaire : Université de Versailles Saint-Quentin-en-Yvelines, 2000. 144 p. 16

SALTON G, MCGILL MJ. Introduction to Modern Information Retrieval. New York : McGraw Hill Book Company, 1983. 448 p. 19, 148

SALTON Gerard. *Automatic Text Processing : The Transformation, Analysis, and Retrieval of Information by Computer*. Massachusetts, USA : Addison-Wesley Publishing Company, 1989. 688p. 19

SEKIMIZU T, PARK H, TSUJII J. Identifying the interaction between genes and gene products based on frequently seen verbs in medline abstracts. *Proccedings of the ninth Workshop on Genome Informatics (GIW'98)*. Universal Academy Press, ISBN 4-946443-52-5, 1998, p. 62-71 38

SHATKAY Hagit, EDWARDS Stephen, WILBUR W. John, BOGUSKI Mark. Genes, Themes and Microarrays - Using Information Retrieval for Large-Scale Gene Analysis. *International Conference on Intelligent Systems for Molecular Biology (ISMB 2000)*, 2000. 35

STAPLEY BJ, BENOIT G. Biobibliometrics : Information retrieval and visualization from co-occurrences of gene names in medline asbtracts. *Proceedings of Pacific Symposium on Biocomputing*, 2000, p. 529-540 35, 149

STEPHENS M, PALAKAL M, MUKHOPADHYAY S, RAJE R, MOSTAFA J. Detecting Gene Relations from MEDLINE Abstracts. *Pacific Symposium on Biocomputing n°6*, 2001, p. 483-496 36

STOESSER G, BAKER W, VAN DEN BROEK A, CAMON E, GARCIA-PASTOR M, KANZ C, KULIKOVA T, LOMBARD V, LOPEZ R, PARKINSON H, REDASCHI N, STERK P, STOEHR P, TULI MA. The EMBL Nucleotide Sequence Database. *Nucleic Acids Res*, 2001, vol. 29, n°1, p. 17-21 30

TANABE L, SCHERF U, SMITH LH, LEE JK, HUNTER L, WEINSTEIN JN. MedMiner : an Internet text-mining tool for biomedical information, with application to gene expression profiling. *BioTechniques*, 1999, vol. 27, p. 1210-1217 26, 34

The chipping forecast. *Nature Genetics*. 1999, supplement, vol. 21, n°1, p. 1-60 16

The genome sequencing consortium. Initial sequencing and analysis of the human genome. *Nature*. 2001, n°409, p. 860-921 15

THOMAS J, MILWARD D, OUZOUNIS C, PULMAN S, CARROLL M. Automatic extraction of protein interactions from scientific abstracts. *Proceedings of the Pac Symp Biocomput*, 2000. p. 541-552 28, 38

TURNER WA, GHERBI R. Hybridation des Savoirs. *Réunion de l'ADEST sur le thème Social Informatics : Outiller les Pratiques Collectives Distribuées*, Paris, 5 juin 2000. On line http://www.upmf-grenoble.fr/adest/seminaires/ 27

USUZAKA S, SIM KL, TANAKA M. A machine learning approach to reducint the work of experts in article selection from database: a case study for regulatory relations of S. cerevisiae genes in MEDLINE. *Proceedings of the Ninth Workshop on Genome Informatics*, 1998, p. 91-101 19

VAN-RIJSBERGEN CJ. *Information Retrieval*. 2nd édition. London, UK : Butterworths, 1979. http://www.dcs.gla.ac.uk/Keith/Preface.html 19

WAIN HM, BLAKE JA, BRUFORD EA, MALTAIS LJ, POVEY S. *Report of ASHG-NW Gene Nomenclature Workshop*. ASHG-NW Gene Nomenclature Workshop. Philadelphie, Octobre 2000. Accessible en ligne <http://www.gene.ucl.ac.uk/nomenclature/ashgnw_report.html> 13

WASINGER VC, CORDWELL SJ, CERPA-POLJAK A, YAN JX, GOOLEY AA, WILKINS MR, DUNCAN MW, HARRIS R, WILLIAMS KL, HUMPHERY-SMITH I. Progress with gene product mapping of the Mollicutes : Mycoplasma genitalium. *Electrophoresis*. 1995, vol. 16, n° 7, p. 1090-1094 16

WILKINS MR, SANCHEZ JC, GOOLEY AA, APPEL RD, HUMPHERY-SMITH I, HOCHSTRASSER DF, WILLIAMS KL. Progress with proteome projects : why all proteins expressed by a genome should be identified and how to do it. *Biotechnol Genet Eng Rev*. 1996, vol. 13, p. 19-50 16

WONG Limsoon. A Protein Interaction Extraction System. *Proceedings of Pacific Symposium on Biocomputing 2001*, p. 520-530. 38

XENARIOS Ioannis, RICE Danny W, SALWINSKI Lukasz, BARON Marisa K, MARCOTTE Edward M, EISENBERG David. DIP : the Database of Interacting Proteins. *Nucleic Acids Research*, 2000, Vol. 28, No. 1, p. 289-291 148

YAMAMOTO Yasunori, OHTA Yoshihiro, UCHIYAMA Ikuo, TAKAGI Toshihisa. Constructing a dictionary of biological terms for information extraction. AKUTSU T, ASAI K, HAGIYA M., KUHARA S, MIYANO S, NAKAI K. *Genome Informatics Workshop*. Tokyo : Universal Academy Press, 1996 30

YOSHIDA Mikio, FUKUDA Ken-ichiro, TAKAGI Toshihisa. PNAD-CSS : a workbench for constructing a protein name abbreviation dictionary. *Bioinformatics*, 2000, vol.16, n° 2, p.169-175 27, 30

YOSHIDA Mikio; FUKUDA Kenichiro, TAKAGI Toshihisa. Automatic Construction of Biological Abbreviation Dictionary from Abstracts of Biomedical Papers. *Proceedings. of the Genome Informatics Workshop 1998*, 1998, N° 9, p. 288-289 30

Nombre de références : 81

PLAN DÉTAILLÉ

ANNEXE

Tableau 76 Les contradictions du dictionnaire.
Les contradictions présentent dans le dictionnaire issu de *Flybase* sont présentés dans ce tableau. Le label intervient dans les définitions de chacun des deux gènes.

Label	Gène 1	Gène 2
ACE1	Amplification-control-element-on-1 (ACE1)	Chorion protein 38 (Cp38)
ACE3	Amplification-control-element-on-3 (ACE3)	Chorion protein 18 (Cp18)
Als	alae sublatae (als)	nicotinic Acetylcholine receptor alpha 96Aa (nAcRa
And	Androcam (And)	dusky (dy)
Ang	angle wing (ang)	anomogenitals (ano)
angle wing	angle wing (ang)	angle winglike (agl)
Ant	antennaless (ant)	empty spiracles (ems)
Apa	Apigmented abdomen (Apa)	Saposin-related (Sap-r)
ARS	Autonomously Replicating Sequence (ARS)	Arylsulfatase (Ars)
Aurora	aurora (aur)	aurora transposable element (aurora-element)
Bald	bald (bld)	balding (bd)
Bam	bag of marbles (bam)	breaks at metaphase (btm)
Bb	Bubble (Bb)	lethal (2) 37Bb (l(2)37Bb)
Bb	bobbed (bb)	Y-bobbed (Ybb)
Bc	Black cells (Bc)	Catecholamines up (Catsup)
Bg	Bag (Bg)	currant bun (cub)
Br	Bridged (Br)	wingless (wg)
Bsh	brain-specific homeobox (bsh)	bushy
Bxd	bxd	Ultrabithorax (Ubx)
Cam	Calmodulin (Cam)	Calcium/calmodulin dependent protein kinase II (Ca
Cap	capon (cap)	Calphotin (Cpn)
Cap	capon (cap)	Chromosome-associated protein (Cap)
Cas	castor (cas)	cashews (cash)
Cat	Catalase (Cat)	Choline acetyltransferase (Cha)
CAT	CAT	Catalase (Cat)
Cb	Curled blistered (Cb)	lethal (2) 37Cb (l(2)37Cb)
Cel	celibate (cel)	cell lethal (cell)
Cf	cleft (cf)	control of female fertility (cff)
Chi	chickadee (chi)	chianti
Chn	charlatan (chn)	Cha's neighbour (Chn)
Chr	chrome (chr)	chrowded (chrw)
Cht	chaetae (cht)	chaste
Cleft	cleft (cf)	water wings (wtw)
Cm	carmine (cm)	crumpled (cmp)
Col	collier (col)	courtless (crl)
Com	compressed (com)	comatose (comt)
Cor	corallium (cor)	coracle (cora)
Cor	corallium (cor)	cortex (cort)
Cr	crisp (cr)	Compensatory Response (CR)
Cre	causes recombination (cre)	mirror (mirr)
Cro	croaker (cro)	pentagon (ptg)
Ctl	coatless (ctl)	cutlet
Curl	Curl (Cu)	Curled 3 (Cu-3)
Curved	curved (c)	curvi (cui)
Cx	curlex (cx)	crossover suppressor on 3 of Gowan (c(3)G)
Cytochrome C	Cytochrome C (Cyt-c2)	Cyt-c1

Label	Gène 1	Gène 2
Da	daughterless (da)	darky (dar)
Dco	discs overgrown (dco)	cAMP-dependent protein kinase 1 (Pka-C1)
Dep	depressed (dep)	defective proventriculus (defective-proventriculus
dl	dorsal (dl)	duplicated legs (dpl)
dm	diminutive (dm)	dimorphos (di)
dos	daughter of sevenless (dos)	dosach (dc)
dr	droopy (dr)	droopy wings (drw)
E(spl)	Enhancer of split (E(spl))	E(spl) region transcript mbeta (HLHmbeta)
E(var)8	Enhancer of variegation 8 (E(var)8)	Enhancer of variegation 3 (E(var)87F)
eag	ether a go-go (eag)	eagle (eg)
early	early (eay)	lodestar (lds)
eb	extra bristles (eb)	sable (s)
ecl	echinus-like (ecl)	eclipse (ecp)
elg	elongatus (elg)	Ets at 97D (Ets97D)
elk	eag-like K + channel (elk)	easterlike (ealk)
Est-9	Esterase 9 (Est-9)	Esterase C (Est-C)
Fab-7	Fab-7	Abdominal B (Abd-B)
fb	fine bristle (fb)	fine bristles (fbr)
flp	flipper (flp)	FLIP recombinase (FLP1)
fltI	flight defective I (fltI)	flightless I (fliI)
Frd	Freckled (Frd)	Frda
fs(1)A59	female sterile (1) A59 (fs(1)A59)	Yolk protein 2 (Yp2)
fs(2)B	female sterile (2) Bridges (fs(2)B)	female sterile (2) 261505 (fs(2)261505)
gd	gastrulation-defective (gd)	giant discs 1 (l(2)gd1)
Gdh	Glutamate dehydrogenase (Gdh)	Glycerol 3 phosphate dehydrogenase (Gpdh)
glide	glide (gli)	glial cells missing (gcm)
gor	gorp (gor)	gorbun (grb)
Got2	Glutamate oxaloacetate transaminase 2 (Got2)	Glutamate oxaloacetate transaminase 1 (Got1)
H1	haemolymph protein 1 (H1)	Histone H1 (His1)
H1	haemolymph protein 1 (H1)	Su(osk)H1
hal	halted (hal)	halley (hall)
hdc	headcase (hdc)	Histidine decarboxylase (Hdc)
hsk	helter-skelter (hsk)	grainy head (grh)
Hsp22	Heat shock protein 22 (Hsp22)	Heat shock protein 23 (Hsp23)
iab-4	iab-4	abdominal A (abd-A)
in	inturned (in)	inflated (if)
Ket	Kettin (Ket)	Ketel (Fs(2)Ket)
kiwi	kiwi	sugarless (sgl)
kn	knot (kn)	knirps (kni)
l(2)Sp9b	lethal (2) Sp9b (l(2)Sp9b)	lethal (2) 41Ae (l(2)41Ae)
ld	loboid (ld)	l(2)k05415
lio	linotte (lio)	limbo
lio	linotte (lio)	derailed (drl)
lme	lame (lme)	lethal (2) meander (l(2)me)
mas	masquerade (mas)	myoblast specific (myoblast-specific)
md	melanotic lesions (md)	minidiscs (mnd)
Mdh2	Malate dehydrogenase 2 (Mdh2)	Malate dehydrogenase 1 (Mdh1)
Me	Moire (Me)	Malic enzyme (Men)
Met	Metatarsi irregular (Met)	Resistance to Juvenile Hormone (Rst(1)JH)
mgt	midget (mgt)	maggot (mgg)
mis	misproportioned (mis)	canoe (cno)
misguided	misguided	longitudinals lacking (lola)
mo	micro-oculus (mo)	moorish (moo)
mrr	myosin rod-related (mrr)	mirror (mirr)
mud	mushroom body defect (mud)	mudlike (mudl)

Label	Gène 1	Gène 2
ne	nicked eye (ne)	negro
neu	neuter (neu)	neuralized (neur)
neur	neuralized (neur)	neuter (neu)
Nuc3	Nuclease 3 (Nuc3)	mutagen-sensitive 308 (mus308)
Or	orange (or)	oculi rugosissimi (oculi-rufosissimi)
Parted	parted (ptd)	abrupt (ab)
Pat	patchytergum (pat)	ptc
Patch	patch	ptc
Pd	purpleoid (pd)	proboscipedia (pb)
Peb	pebbled (peb)	Protein ejaculatory bulb (Peb)
pebbled	pebbled (peb)	rugose (rg)
Pen	penguin (pen)	Pendulin (Pen)
Pen	penguin (pen)	pendolino (peo)
penguin	penguin (pen)	pygoscelis (pyg)
Pg	prong (pg)	pigmy (pig)
Phm	phantom (phm)	polyhomeotic proximal (ph-p)
Phm	phantom (phm)	polyhomeotic distal (ph-d)
Pio	piopio (pio)	linotte (lio)
PKA	Protein Kinase A (PKA)	Pka-R1
PKA	Protein Kinase A (PKA)	cAMP-dependent protein kinase 1 (Pka-C1)
pl	pleated (pl)	pallid (pld)
poc	polycephalon (poc)	porcupine (por)
psb	pseudobeadex (psb)	postscutellar bristles (pbr)
psc	pseudoscute (psc)	Posterior sex combs (Psc)
ptd	parted (ptd)	abrupt (ab)
ptl	pointless (ptl)	pathless (ptls)
pun	puny (pun)	punt (put)
RacA	RacA	Rac1
rag	ragged (rag)	rags
Rap1	repressor/activator protein 1 (Rap1)	Roughened (R)
ras	raspberry (ras)	Ras oncogene at 64B (Ras64B)
ras	raspberry (ras)	Ras oncogene at 85D (Ras85D)
raven	raven (rv)	ravenoid (rvn)
rdb	reddish brown (rdb)	reduced bristles (rbl)
rdp	reduplicated (rdp)	broad (br)
re	reduced eyes (re)	rough eye (rey)
rea	rearranged tergites (rea)	anon-92Ed
rem	remnants (rem)	mushroom body miniature B (mbmB)
ret	reticulated (ret)	reticent (reti)
rh	roughish (rh)	glass (gl)
rm	rimy (rm)	rumpled (rmp)
rol	reduced optic lobes (rol)	rho-like (RhoL)
RpL11	RpL11	Ribosomal protein S9 (RpS9)
rsd	raised (rsd)	Actin 88F (Act88F)
rub	rubroad (rub)	rubbish (rubb)
rumpled	rumpled (rmp)	pineapple eye (pie)
rw	raised wing (rw)	red wine (rwi)
sa	spermatocyte arrest (sa)	cramped (crm)
sad	shadow (sad)	nicotinic Acetylcholine receptor alpha 96Ab (nAcRa
sam	sperm amotile (sam)	sallimus (sls)
sat	satin (sat)	fruitless (fru)
sb	soft brown (sb)	minute-like (ml)
Sc	Scotched eye (Sc)	Scoop (Scp)
scd	sex combs distal (scd)	scattered
sd	scalloped (sd)	spread (sprd)

Label	Gène 1	Gène 2
Sex combs extra	Sex combs extra (Sce)	Antennapedia (Antp)
shd	shade (shd)	Notch (N)
shm	short macros (shm)	single-minded (sim)
short wing	short wing (sw)	short winged (sh)
shv	shiva (shv)	decapentaplegic (dpp)
shv	shiva (shv)	shortened veins (svs)
sl	small wing (sl)	sluggish A (slgA)
sm	smooth (sm)	smoky (smk)
smo	smoothened (smo)	smooth (sm)
smooth	smooth (sm)	smoothened (smo)
snb	snowballs (snb)	pink (p)
spa	sparkling (spa)	Serine pyruvate aminotransferase (Spat)
spg	sponge (spg)	spaghetti squash (sqh)
spx	split thorax (spx)	spreadex (sdx)
Srf	Surf wings (Srf)	blistered (bs)
stb	short bristle (stb)	starburst (strb)
stb	short bristle (stb)	strabismus (stbm)
std	staroid (std)	Serrate (Ser)
Ste	Stellate (Ste)	Suppressor of Stellate (Su(Ste))
stk	sticking (stk)	steamer duck (stck)
str	stripes (str)	thickveins (tkv)
struthio	struthio (stru)	held out wings (how)
sty	sprouty (sty)	giant lens (gil)
Su(var)	Suppressor of variegation (Su(var))	Suppressor of variegation 2-1 (Su(var)2-1)
Suppressor of variegation	Suppressor of variegation (Su(var))	Protein phosphatase 1 at 87B (Pp1-87B)
swa	swallow (swa)	cramped (crm)
ta	tapered (ta)	tarry (tar)
tb	tiny bristle (tb)	tracheae broken (tbr)
te	tenerchaetae (te)	tete
ter	terraced (ter)	terminus (term)
thi	thickhead (thi)	thickened veins (thiv)
tk	thick (tk)	TK
Tm1	Tropomyosin 1 (Tm1)	Tropomyosin 2 (Tm2)
ton	tonochaetae (ton)	tondo
Tre	Trehalose-sensitivity (Tre)	Trehalase (Treh)
Trf	TBP-related factor (Trf)	Transferrin
tw	twisted (tw)	twins (tws)
twe	twine (twe)	tweety (tty)
unp	unexpanded (unp)	unplugged (unpg)
Vinculin	Vinculin (Vin)	Vinculin at 2EF (Vin2EF)
wap	wings apart (wap)	wings apart-like (wapl)
Wrinkled	Wrinkled (W)	Wrinkle (Wr)
Z	Zerknittert (Z)	anon-84Ba
zip	zipper (zip)	unzipped (uzip)
zipper	zipper (zip)	unzipped (uzip)

Tableau 77 Liste des labels de type de reconnaissance *mots vides si début de phrase*
Les labels de cette catégorie sont très ambigus s'ils sont placés en première position dans une phrase.

An	And	As	At	Be	By	Can	Co	Did	Do	For	Had
Her	How	If	In	Is	Low	Me	None	Not	Off	On	Or
Per	Re	She	So	To	Up	Us	Ve	We	Who	With	

Tableau 78 Liste des labels de type de reconnaissance *mots vides*
Ces labels sont très ambigus quelles que soient leurs positions dans la phase.

an	and	as	at	be	by	can	co	did
do	for	had	her	how	if	in	is	low
me	no	none	not	off	on	or	per	she
so	to	up	us	we	who	with		

Tableau 79 Labels de type de reconnaissance *ambigu en début de phrase*
Ces labels sont ambigus quand ils sont placés en début de phrase mais ne sont pas des *mots vides*.

Abdominal	Abrupt	Adipose	Amalgam	Amber	Ambiguous
Antenna	Arm	Arrest	Bag	Band	Bent
Blocked	Blood	Blunt	Bordered	Brief	Bristle
Broad	Cap	Cardinal	Clipped	Condensed	Cortex
Dark	Defective	Deformed	Dense	Depleted	Depressed
Disrupted	Divergent	Dorsal	Double	Drop	Early
Eg	Erratic	Expanded	Extended	Eye	Furrow
Giant	High	Inactive	Juvenile	Labial	Large
Leg	Limited	Lines	Malformed	Map	Mid
Midline	Midway	Miniature	Minute	Missing	Morula
Multiple	Narrow	Oblique	Opaque	Open	Paired
Pale	Period	Pointed	Pre	Raised	Ray
Reduced	Retained	Separated	Shifted	Silver	Similar
Small	Smaller	Spliced	Spread	Sticky	Streak
Stripe	Stripes	Syndrome	Ten	Terminus	Thick
Thin	Thread	Tilt	Tiny	Trunk	Tumor
Twisted	Uneven	Unfolded	Vein	Weak	Abbreviated
Approximated	Attenuated	Compressed	Daughterless	Ectodermal	Naked cuticle
Ecdysone receptor		Rudimentary	Uncoordinated	Reversed polarity	

Tableau 80 Label de type de reconnaissance *terme spécifique*
Ces labels ne sont pas des noms de gène mais des termes inclus des noms de gènes.

animal cap	Cell-cell	chromosome arm	cn bw chromosome	
C-terminal arm	disrupted polarity	dorsal cell	dorsal cells	dorsal closure
dorsal ectoderm	dorsal epidermis	dorsal fate	dorsal follicle	dorsal half
dorsal midline	dorsal or ventral	dorsal pattern	Dorsal side	dorsal side
dorsal vessel	dorsal-side	dorsal-specific	dorsal-ventral	e.g.
entire disc	G-phase	i.e.	imaginal disc	mis- expression
N-terminal	N-terminal arm	P element mediated		patch of
P- element-mediated		P-element-mediated		pupal stage
pupal stages	ring canal	ring canals	S phase	see ref
slight effect	S-phase	ventral furrow	ventral side	wing disc
adaptive response (AR)		morphogenetic furrow		
P element transformation		P-element transformation		

Tableau 81 Label de type de reconnaissance *plutôt ambigu*
Ces labels ont été jugés à priori ambigu.

abdominal	act	al	ambiguous	antenna	arrest	attenuated
bag	band	blocked	blood	bp	brief	bristle
broad	c	C	cap	Cell	clock	cortex
dark	defective	dense	depleted	depressed	disrupted	divergent
double	drop	early	ectodermal	eg	extended	eye
H	h	high	inactive	juvenile	labial	large
leg	light	limited	lines	ll	m	M
M2	map	mid	midline	midway	minute	mis
missing	MR	multiple	n	N	narrow	old
open	paired	period	pre	R	r	raised
ras	ray	re	reduced	retained	s	scattered
SD	separated	set	shifted	similar	small	smaller
spliced	spread	stripe	stripes	T	terminus	trunk
tumor	tumorous	twisted	ve	VE	vein	weak
condensed	deformed	malformed	rudimentary	naked cuticle		
ovarian tumor		reversed polarity				

Tableau 82 Labels de type de reconnaissance *peut-être ambigu*
Ces labels ont été jugés à priori comme possiblement ambigu.

abbreviated	ABRUPT	abrupt	adipose	amalgam	amber	approximated	bent
blunt	bordered	cardinal	CELL	clipped	displaced	DORSAL	erratic
expanded	FURROW		MAP	miniature	morula	oblique	opaque
OPEN	pale	PERIOD	pointed	PRE	sd	silver	SMALL
SMALLER	sticky	streak	thread	tilt	tiny	uncoordinated	
uneven	unfolded	VEIN	compressed		syndrome		

Tableau 83 Labels de type de reconnaissance *désambiguïsation en cours*
Ces labels ont déjà reçu des termes de désambiguïsation mais il n'y en a pas encore suffisamment pour que ce soit satisfaisant.

cell	furrow	Ring	ring
S	side	Side	

Tableau 84 Label de divers type de reconnaissance
Certain type de reconnaissance compte peu de membres. Nous les donnons donc dans un tableau unique.

Trop ambigu	Mr	NO	Y
Désambigüé	arm	dorsal	
Ambiguïté constatée mais marginale	Cdi	CS	ME

Tableau 85 Labels de type de reconnaissance *peu ambigus*

Ces labels ne sont pas ambigus pour la plupart car il se distingue de termes anglais courant par la casse.

ABDOMINAL	ACT	Act	ADIPOSE	AL	Al
AMALGAM	AMBER	AMBIGUOUS	AN	AND	ANTENNA
APPROXIMATED		ARM	ARREST	AS	AT
BAG	BAND	BE	BENT	BLOCKED	BLOOD
BLUNT	BORDERED	BP	Bp	BRIEF	BRISTLE
BROAD	BY	CAN	CAP	CARDINAL	CLIPPED
CO	COMPRESSED		CONDENSED	CORTEX	DARK
DEFECTIVE	DEFORMED	DENSE	DEPLETED	DEPRESSED	DID
Displaced	DISPLACED	DIVERGENT	DO	DOUBLE	DROP
EARLY	ECDYSONE RECEPTOR		ecdysone receptor		EG
ERRATIC	EXPANDED	EXTENDED	EYE	FOR	giant
GIANT	HAD	HEDGEHOG	hedgehog	Hedgehog	HER
HIGH	HOW	IF	IN	INACTIVE	IS
JUVENILE	LABIAL	LARGE	LEG	LIMITED	LINES
Ll	LL	LOW	Ltd	ltd	LTD
MALFORMED	MID	MIDLINE	MIDWAY	MINIATURE	MINUTE
MISSING	MORULA	mr	MULTIPLE	NARROW	NONE
NOT	OBLIQUE	OFF	ON	OPAQUE	OR
PAIRED	PALE	PER	POINTED	RAISED	RAY
RE	REDUCED	RETAINED	REVERSED POLARITY		RING
Rough	ROUGH	rough	Sd	SEPARATED	SHE
SHIFTED	SIDE	SILVER	SIMILAR	SO	SPLICED
SPREAD	STICKY	stranded	Stranded	STRANDED	STREAK
STRIPE	STRIPES	suffix	Suffix	SUFFIX	SYNDROME
TERMINUS	thick	THICK	THIN	thin	THREAD
TILT	TINY	TO	TRUNK	Tube	TUBE
tube	TUMOR	TWISTED	UNCOORDINATED		UNEVEN
UNFOLDED	UP	US	WE	WEAK	WHO
WITH	ABBREVIATED		ATTENUATED		RUDIMENTARY
DAUGHTERLESS		ECTODERMAL		NAKED CUTICLE	

Tableau 86 Label de type de reconnaissance *spécifié univoque*

Ces labels, pour la plupart saisie par l'annotateur, ont été jugés parfaitement univoques.

abdominal- A		arrestins	Brachyury	phosrestins I		Dorsal-Cactus	Pc-group
E2F-1	E2F-2	E2F-3	D-mekts	extra-macrochaete		msl	Musca PRI
Musca PRIs		human E2F	Runt	Torso	white-apricot	white-blood	
adaptive response			dorsal switch protein		mammalian transcription factor Sp1		
human proto-oncogene pbx1			Antennapedia and bithorax complexes			dorsal-group	

Tableau 87 Données du graphique de la figure 6

Calcul des taux de rappel et de précision pour les occurrences d'interactions (processus *2RDG*)

Seuil	Automatique	Expert	Confirmée	Rappel (%)	Précision (%)
-2	225	89	80	90	36
-0,25	199	89	79	89	40
-0,2	184	89	76	85	41
-0,19	180	89	75	84	42
-0,18	179	89	75	84	42
-0,17	176	89	74	83	42
-0,16	172	89	73	82	42
-0,15	169	89	72	81	43
-0,14	164	89	72	81	44
-0,13	161	89	70	79	43
-0,12	156	89	67	75	43
-0,11	153	89	65	73	42
-0,1	148	89	65	73	44
-0,09	143	89	64	72	45
-0,08	139	89	63	71	45
-0,07	136	89	61	69	45
-0,06	128	89	60	67	47
-0,05	125	89	58	65	46
-0,04	117	89	54	61	46
-0,03	110	89	52	58	47
-0,02	103	89	49	55	48
-0,01	100	89	48	54	48
0	90	89	43	48	48
0,01	85	89	42	47	49
0,02	79	89	38	43	48
0,03	76	89	37	42	49
0,04	72	89	35	39	49
0,05	68	89	35	39	51
0,06	65	89	33	37	51
0,07	57	89	29	33	51
0,08	55	89	29	33	53
0,09	51	89	29	33	57
0,1	49	89	29	33	59
0,11	46	89	28	31	61
0,12	44	89	28	31	64
0,13	39	89	24	27	62
0,14	37	89	23	26	62
0,15	33	89	20	22	61
0,16	33	89	20	22	61
0,17	30	89	18	20	60
0,18	29	89	17	19	59
0,19	27	89	16	18	59
0,2	26	89	15	17	58
0,21	24	89	15	17	63
0,22	21	89	14	16	67
0,23	18	89	12	13	67
0,24	16	89	12	13	75
0,25	13	89	10	11	77

Tableau 88 Données du grapique de la figure 7 et de la figure 9
Concerne le processus *2RDG* et *nRDG*

Seuil	Automat.	Expert	Confirmée	Rappel (%)	Précision (%)	Rappel *nRDG*
-2	134	55	52	95	39	34
-0,2	115	55	51	93	44	33
-0,15	109	55	50	91	46	32
-0,1	99	55	49	89	49	32
-0,09	95	55	48	87	51	31
-0,08	94	55	47	85	50	31
-0,07	91	55	45	82	49	29
-0,06	88	55	45	82	51	29
-0,05	86	55	44	80	51	29
-0,04	84	55	43	78	51	28
-0,03	81	55	42	76	52	27
-0,02	78	55	41	75	53	27
-0,01	76	55	40	73	53	26
0	69	55	36	65	52	23
0,01	65	55	35	64	54	23
0,02	61	55	32	58	52	21
0,03	58	55	31	56	53	20
0,04	56	55	29	53	52	19
0,05	54	55	29	53	54	19
0,06	52	55	28	51	54	18
0,07	46	55	26	47	57	17
0,08	44	55	26	47	59	17
0,09	41	55	26	47	63	17
0,1	40	55	25	45	63	16
0,11	37	55	24	44	65	16
0,12	36	55	23	42	64	15
0,13	32	55	20	36	63	13
0,14	30	55	19	35	63	12
0,15	27	55	16	29	59	10
0,16	27	55	16	29	59	10
0,17	25	55	15	27	60	10
0,18	24	55	14	25	58	9
0,19	22	55	13	24	59	8
0,2	21	55	12	22	57	8
0,21	20	55	12	22	60	8
0,22	17	55	11	20	65	7
0,23	16	55	10	18	63	6
0,24	14	55	10	18	71	6

Tableau 89 Données de la figure 8
Concerne le processus *nRDG*

Seuil	Automatique	Expert	Confirmée	Rappel (%)	Précision (%)
-2	986	154	145	94	15
-0,2	885	154	144	94	16
-0,15	792	154	137	89	17
-0,14	769	154	136	88	18
-0,13	742	154	133	86	18
-0,12	732	154	133	86	18
-0,11	725	154	133	86	18
-0,1	707	154	131	85	19
-0,09	689	154	130	84	19
-0,08	657	154	128	83	19
-0,07	651	154	127	82	20
-0,06	643	154	127	82	20
-0,05	602	154	121	79	20
-0,04	577	154	120	78	21
-0,03	565	154	119	77	21
-0,02	551	154	117	76	21
-0,01	534	154	116	75	22
0	503	154	110	71	22
0,01	480	154	106	69	22
0,02	457	154	104	68	23
0,03	449	154	101	66	22
0,04	421	154	89	58	21
0,05	406	154	88	57	22
0,06	373	154	86	56	23
0,07	359	154	83	54	23
0,08	344	154	78	51	23
0,09	292	154	70	45	24
0,1	290	154	70	45	24
0,11	267	154	64	42	24
0,12	257	154	63	41	25
0,13	227	154	58	38	26
0,14	216	154	55	36	25
0,15	208	154	51	33	25
0,16	178	154	51	33	29
0,17	151	154	46	30	30
0,18	144	154	44	29	31
0,19	139	154	43	28	31
0,2	110	154	31	20	28
0,21	101	154	28	18	28
0,22	94	154	26	17	28

Seuil	Automatique	Expert	Confirmée	Rappel (%)	Précision (%)
0,23	84	154	22	14	26
0,24	77	154	22	14	29
0,25	63	154	19	12	30
0,26	36	154	13	8	36
0,27	35	154	12	8	34
0,28	31	154	11	7	35
0,29	29	154	10	6	34
0,3	28	154	10	6	36

Tableau 90 Performance du processus *2G*

Ce sont les données utilisées pour la figure 12.

Seuil	Automatique	Expert	Confirmée	Rappel (%)	Précision (%)	Rappel *nG* (%)
-2	95	62	58	94	61	41
-0,3	93	62	58	94	62	41
-0,2	85	62	57	92	67	40
-0,15	80	62	56	90	70	40
-0,1	75	62	55	89	73	39
-0,09	73	62	55	89	75	39
-0,08	71	62	54	87	76	38
-0,07	70	62	53	85	76	38
-0,06	69	62	53	85	77	38
-0,05	67	62	52	84	78	37
-0,04	65	62	51	82	78	36
-0,03	63	62	49	79	78	35
-0,02	62	62	48	77	77	34
-0,01	61	62	47	76	77	33
0	56	62	43	69	77	30
0,01	53	62	42	68	79	30
0,02	50	62	39	63	78	28
0,03	47	62	37	60	79	26
0,04	45	62	35	56	78	25
0,05	42	62	34	55	81	24
0,06	41	62	33	53	80	23
0,07	38	62	31	50	82	22
0,08	37	62	31	50	84	22
0,09	35	62	31	50	89	22
0,1	35	62	31	50	89	22
0,11	33	62	29	47	88	21
0,12	33	62	29	47	88	21
0,13	30	62	26	42	87	18
0,14	28	62	24	39	86	17
0,15	24	62	20	32	83	14
0,16	24	62	20	32	83	14
0,17	21	62	18	29	86	13
0,18	19	62	16	26	84	11
0,19	17	62	15	24	88	11
0,2	15	62	13	21	87	9
0,21	15	62	13	21	87	9
0,22	14	62	12	19	86	9
0,23	13	62	11	18	85	8
0,24	13	62	11	18	85	8

Tableau 91 Données du calcul rappel-précision pour *nG*

Ce tableau fournis les données utilisées pour la figure 11.

Seuil	Automatique	Expert	Confirmée	Rappel (%)	Précision (%)
-2	282	141	133	94	47
-0,2	276	141	132	94	48
-0,2	269	141	130	92	48
-0,2	268	141	130	92	49
-0,2	265	141	128	91	48
-0,2	261	141	125	89	48
-0,2	260	141	125	89	48
-0,1	257	141	124	88	48
-0,1	251	141	121	86	48
-0,1	250	141	121	86	48
-0,1	250	141	121	86	48
-0,1	239	141	119	84	50
-0,1	235	141	118	84	50
-0,1	207	141	117	83	57
-0,1	206	141	116	82	56
-0,1	205	141	116	82	57

Seuil	Automatique	Expert	Confirmée	Rappel (%)	Précision (%)
-0,1	193	141	110	78	57
-0	188	141	109	77	58
-0	187	141	108	77	58
-0	183	141	106	75	58
-0	182	141	105	74	58
0	170	141	99	70	58
0,01	158	141	95	67	60
0,02	156	141	93	66	60
0,03	154	141	91	65	59
0,04	142	141	79	56	56
0,05	139	141	78	55	56
0,06	133	141	76	54	57
0,07	130	141	74	52	57
0,08	121	141	68	48	56
0,09	108	141	64	45	59
0,1	108	141	64	45	59
0,11	91	141	59	42	65
0,12	88	141	58	41	66
0,13	80	141	54	38	68
0,14	76	141	51	36	67
0,15	72	141	47	33	65
0,16	72	141	47	33	65
0,17	63	141	42	30	67
0,18	58	141	40	28	69
0,19	57	141	39	28	68
0,2	36	141	27	19	75
0,21	33	141	24	17	73
0,22	29	141	22	16	76
0,23	25	141	18	13	72
0,24	25	141	18	13	72
0,25	21	141	15	11	71
0,26	13	141	10	7	77
0,27	13	141	10	7	77

Tableau 92 Données pour les interactions répétées issues du processus *2G*.
Ce sont les données de la séries *2Gr* de la figure 13.

Seuil	Automatique	Expert	Confirmée	Rappel (%)	Précision (%)
-2	39	62	27	44	69
-0,25	32	62	26	42	81
-0,15	30	62	25	40	83
-0,14	30	62	25	40	83
-0,13	30	62	25	40	83
-0,12	29	62	24	39	83
-0,11	28	62	23	37	82
-0,1	28	62	23	37	82
-0,09	25	62	21	34	84
-0,08	25	62	21	34	84
-0,07	24	62	20	32	83
-0,06	21	62	18	29	86
-0,05	20	62	17	27	85
-0,04	18	62	15	24	83
-0,03	18	62	15	24	83
-0,02	16	62	14	23	88
-0,01	14	62	12	19	86
0	12	62	11	18	92
0,01	11	62	10	16	91
0,02	11	62	10	16	91
0,03	11	62	10	16	91

Tableau 93 Données pour les interactions répétées issues du processus *nG*

Ce sont les données de la série *nGr* de la figure 14.

Seuil	Automatique	Expert	Confirmée	Rappel (%)	Précision (%)
-2	89	141	52	37	58
-0,25	80	141	51	36	64
-0,15	74	141	50	35	68
-0,1	71	141	48	34	68
-0,09	70	141	48	34	69
-0,08	68	141	47	33	69
-0,07	67	141	46	33	69
-0,06	64	141	44	31	69
-0,05	60	141	42	30	70
-0,04	55	141	38	27	69
-0,03	54	141	37	26	69
-0,02	52	141	36	26	69
-0,01	50	141	34	24	68
0	46	141	31	22	67
0,01	45	141	30	21	67
0,02	44	141	29	21	66
0,03	44	141	29	21	66
0,04	37	141	25	18	68
0,05	37	141	25	18	68
0,06	36	141	24	17	67
0,07	36	141	24	17	67
0,08	32	141	24	17	75
0,09	25	141	21	15	84
0,1	24	141	20	14	83
0,11	19	141	17	12	89
0,12	16	141	15	11	94
0,13	13	141	12	9	92
0,14	13	141	12	9	92
0,15	12	141	11	8	92

Title : Experiment in factual databases and bibliographical databases integration:
Gene identification in Medline from Flybase description and application in information extraction about genetics or molecular interaction in publications.

Abstract : Databases have become an essential working tool for research in genetics. Factual databases organize the knowledge accumulated in electronics encyclopedias and in experimental-results databanks. Bibliographical databases give access to the most precise and comprehensive information. It is necessary to completely couple these two types of databases to allow automatic interaction.
The question is either to document factual databases with bibliographical references, or to extract information directly from scientific publications.
To explain clearly the difficulty of working between the two types and also offer a solution, we took the genes, and their interactions, of the Drosophila as a study case. The genetic interactions are essential phenomena in understanding the way the genes collaborate in one function. We chose the model genetic organism Drosophila because its genes are well described in the electronic encyclopedia Flybase and their interactions are well described in the bibliographic database Medline.
First, we built a system that makes it possible to create links between Flybase and Medline. These links consist in documenting every gene described in Flybase by Medline bibliographical references. It is thus a question of identifying the genes in Medline's summaries. This task is difficult to undertake automatically because of the complexity of the naming (existence of alias', of abbreviations and vague terms, composition of terms using names of genes, etc.) and of the possible confusion between some gene names and some words of the common vocabulary such as abdominal, labial, early, N, etc.
In the second instance, our work consisted of establishing a list of likely interactions from a set of Medline's summaries. We have done this work by statistical analysis of the vocabulary in the summaries.

Keywords : Information extraction, information retrieval, natural language processing, text processing, terminology, nomenclature, genes, genetics interaction, molecular interaction, bioinformatics, Drosophila melanogaster, Medline, Flybase

Résumé :La thèse propose des solutions pour mettre automatiquement en relation des informations bibliographiques avec des informations factuelles. Les bases de données bibliographiques donne accès à l'information la plus exhaustive et la plus précise tandis que les bases de données factuelles organisent le savoir accumulé dans des encyclopédies électronique ou dans des banques de résultats d'expériences. Coupler ces deux types de bases de données est nécessaire. Il s'agit soit de documenter des bases de données factuelles avec des références bibliographiques, soit d'extraire de l'information directement à partir de la littérature scientifique.
Nous avons pris l'exemple des gènes et de leurs interactions chez la Drosophile. La Drosophile est un organisme modèle en génétique et l'analyse des interactions génétiques ou moléculaires permet de comprendre comment plusieurs gènes collaborent à une même fonction.
Dans un premier temps, nous avons construit un système qui permet de créer des liens entre Flybase et Medline. Flybase est une encyclopédie électronique sur la Drosophile. Medline est la plus grande base de données bibliographiques dans le domaine des sciences de la vie. Ces liens consistent à identifier dans Medline des gènes décrit dans Flybase. Cette tâche est difficile à automatiser en raison de la complexité de la nomenclature (existence d'alias, d'abréviations et de termes vagues, composition de termes utilisant des noms de gènes, etc.) et de la confusion possible entre certains noms de gènes et des mots du vocabulaire courant.
Dans un second temps, notre travail a consisté à établir une liste d'interactions probables à partir d'un ensemble de résumés issus de Medline. Cela a été fait par l'analyse statistique du vocabulaire utilisé.
La méthode a été testée avec succès et le détail de la mise en œuvre est donné dans le document.

Mots clefs : Extraction d'information, informatique documentaire, statistique textuelle, couplage de bases de données, terminologie, nomenclature, gènes, interaction génétique, interaction moléculaire, bioinformatique, Drosophila Melanogaster, Medline, Flybase.

Résumé en anglais : voir page précédente

Discipline : Science de l'information et de la communication

Laboratoire : CRRM (Case 161)
Centre scientifique de saint Jérôme
13397 Marseille Cedex 20