Index des publications VSST'98

Analyse de données et analyse de mots associés, comparaison d'algorithmes différents sur un corpus concernant la prise en compte du risque dans le developpement des OGM dans le domaine des végétaux M-A De Looze, A Roy, M Reinert, O Jouve, R Coronini

Un outil pour la veille: le maillage technologique de références bibliographiques scientifiques P.Faucompre, L.Quoniam

Enseignement tiré d'un programme de compréhension automatique de comptes rendus médico-techniques M.Roux, V.Ledoray

Système 3AD: un outil de classification à caractère linguistico-mathématique I.Timimi

Les processus d'apprentissage au coeur de la veille technologique dans un environnement de R&D C.Rondeau

Organisation et gestion des connaissances en veille scientifique et technologique X.Polanco, C.Francois, J.Royaute, L.Grivel, D.Besagni, M.Dejean, C.Oto

Création d'hypertextes automatiques appliqués à la veille V.Leveille, H.Rostaing, L.Quoniam

Une méthode de détection des signaux faibles: application à l'émergence des dendrimères C.Roux, B.Dousset

EXTRACTION AUTOMATIQUE ET REPRESENTATION GRAPHIQUE DE DONNEES BIOLOGIQUES : LES INTERACTIONS GENETIQUES ET MOLECULAIRES.
APPLICATION A UN ORGANISME MODELE ET AU GENOME HUMAIN.

Violaine PILLET*, Doctorant deuxième année
Bader ROUDANI*, Doctorant deuxième année
Luc QUONIAM*, Maître de Conférences au CRRM
Bernard JACQ**, Chargé de Recherche au CNRS
*Centre de Recherche Rétrospective de Marseille
Université Aix-Marseille III
FR 13397 Marseille Cedex 20
Tél. (33) 04 91 28 87 40 Fax (33) 04 91 28 87 12
e-mail : violaine@crrm.univ-mrs.fr, luc@crrm.univ-mrs.fr
**Laboratoire de Génétique et Physiologie du Développement - CNRS
Université de Luminy Case 907
FR 13288 Marseille Cedex 9
Tél. (33) 04 91 26 96 00 Fax (33) 04 91 82 06 82
e-mail : jacq@ibdm.univ-mrs.fr

Résumé : Dans les domaines de la recherche et de l'industrie il ne s'agit plus seulement de synthétiser l'information, mais il devient nécessaire de rassembler les informations existantes et de créer des liens entre les différentes données afin de faire émerger de nouvelles connaissances indispensables à la recherche scientifique. Il est donc indispensable de mettre à la disposition des chercheurs, des outils et des méthodes permettant d'extraire de façon automatique, à partir de grands volumes de données (plusieurs dizaines de milliers de références bibliographiques), des données spécifiques, de les traiter puis de les représenter.

En génétique, de nombreuses bases de données décrivent la séquence ou la structure des différents gènes, ARN et protéines mais il n'existe aucune base de données dédiées à l'étude des interactions génétiques et moléculaires et des réseaux qu'elles constituent. Aucune image d'un réseau de régulations et de ses composants ne peut être actuellement déduit des bases de données génétiques et moléculaires existantes. La description et la représentation formelle des réseaux d'interactions et leur assemblage en réseaux régulateurs sont pourtant des étapes essentielles à la compréhension des processus complexes contrôlés par ces gènes. Pour combler ce manque, nous nous proposons d'élaborer une stratégie qui recense les informations existantes sur les interactions, de décrire et représenter ces connaissances puis de créer de nouvelles bases de données structurées, centrées uniquement sur les interactions de plusieurs organismes modèles et pour une partie du génome humain.

Mots-clés : Traitement automatique, réseau, interaction, gène, protéine, Drosophila melanogaster, génome humain, cartographie, bases de données

INTRODUCTION

L'un des enjeux de l'industrie pharmaceutique et de la recherche médicale est de proposer des techniques, thérapies et médicaments qui permettent d'améliorer l'état de santé d'un individu ou mieux, de le guérir, ceci en faisant de plus en plus souvent appel aux nouvelles connaissances génétiques. Pour cela, il est essentiel de comprendre comment fonctionnent les mécanismes génétiques mis en jeu lors du fonctionnement normal des systèmes biologiques, mais aussi et surtout lors des dysfonctionnements provoquant des maladies telles que le cancer ou encore la myopathie.
Il existe de nombreuses sources d'informations sur la structure et la fonction des gènes et des protéines impliqués dans ces maladies cependant il est actuellement très difficile de mettre en relation toutes ces informations et d'en obtenir une vision synthétique. Pour cela les centres de recherches et les industries ont de plus en plus besoin de systèmes d'aide à l'extraction, l'analyse et la représentation synthétique de ces informations. C'est pourquoi, nous nous proposons de définir et mettre en place, à partir d'outils informatiques spécifiques, différentes stratégies afin d'extraire de façon semi-automatique, à partir de sources d'informations déjà existantes, des informations sur les interactions génétiques et moléculaires. Ensuite il est nécessaire de rassembler ces informations, de les traiter puis de les analyser dans le but de faire émerger et valoriser de nouvelles connaissances qui seront stockées dans une base de connaissance structurée, dédiées aux interactions. Cette base sera accessible à la communauté scientifique par le biais des nouveaux réseaux de communications tel que l'Internet.
Au niveau biologique, le recensement et la synthèse des données sur les interactions permettront d'apporter une nouvelle vision ne se concentrant plus sur un gène unique impliqué potentiellement dans une pathologie mais au contraire, sur un ensemble de gènes dont on cherche à décrire et comprendre les interactions physiologiques coordonnées.
Nous montrerons dans cet exposé quelles sont les sources d'informations qui ont été choisies, quelles techniques ont été employées pour établir cette méthodologie et enfin comment représenter les nouvelles connaissances.

EXTRACTION ET REFORMATAGE DES DONNEES

Les sources d'informations

La première phase de l'étude consiste à répertorier un maximum de données sur les interactions génétiques et moléculaires chez un organisme modèle : la drosophile. Le choix de cet organisme est du au fait que c'est celui dont la génétique est la mieux connue et dont de nombreuses données expérimentales ont été publiées et répertoriées dans les bases de données sur cet organisme.
Dans un premier temps, il s'agit de répertorier les sources d'informations électroniques susceptibles de contenir des informations de nature fonctionnelle sur les gènes et les protéines de la drosophile. Une recherche sur le Web à partir de la Virtual Library a permis de détecter la base de données FlyBase [FlyBase 1996]. Cette base de données fondée en 1992 par le National Institute of Health (USA) et le Medical Research Concil (UK) à partir du catalogue de Dan Lindsley et Georginna Zimm : « The Genome of Drosophila melanogaster » [Lindsley 1992], contient actuellement plus de 89000 références de publications sur la drosophile. Elle répertorie plus de 15000 gènes et plus de 51500 allèles. Une étude approfondie de cette base de données a montré que celle-ci est relativement peu structurée et qu'elle n'a pas été créée

pour traiter spécifiquement des interactions entre ADN-protéines, protéines-protéines et ARN-protéines, mais qu'elle contient de façon désordonnée de nombreuses informations relatives aux interactions.

Une étude de la structure de la base de données a été faite. Celle-ci montre qu'elle est constituée d'entrées. Chaque entrée correspond à un gène donné. Chaque gène est décrit dans différents champs (fig.1). L'étude du contenu des différents champs de la base a fait apparaître que seuls deux ou trois champs de la base sont susceptibles de contenir les informations relatives aux interactions moléculaires recherchées. Mais la manière dont sont décrites les interactions n'est pas homogène, c'est à dire qu'il existe différentes formes d'écriture pour décrire une interaction. Il ne sera donc pas facile d'extraire ces données ni de mettre en évidence des termes spécifiques aux interactions.

Il s'agit donc de définir et mettre en place, à partir d'outils informatiques spécifiques, différentes stratégies pour extraire de façon semi-automatique, à partir de FlyBase, toutes les données scientifiques sur les interactions génétiques et moléculaires.

Acquisition semi-automatique des données.

Une fois la base de données télédéchargée (gratuitement à partir du Web) et les informations pertinentes repérées dans les différents champs de la base, une série d'étapes de reformatage ont été effectuées à l'aide du logiciel Infotrans [Infotrans], pour créer une sous-base contenant les données stratégiques.
Les étapes de reformatage consistent tout d'abord à détecter et extraire les différents champs intéressants dont des mini-résumés d'articles puis à restructurer ces champs. Ensuite, un premier filtre est appliqué pour éliminer toutes les références ne contenant pas de mini-résumés. Enfin, une phase de dédoublonnage permet d'éliminer les doublons présents dans la sous-base. Une nouvelle base de données structurée est ainsi obtenue. Celle-ci contient environ 20300 entrées et répertorie 9550 gènes.

Reconnaissance et balisage des noms de gènes et de protéines

A ce stade, le fichier présente encore de nombreuses phrases qui ne traitent pas d'interactions géniques. Il s'agit donc de trouver de nouveaux filtres, qui permettent de supprimer de façon le plus automatique possible, les informations pertinentes. En biologie, lorsque l'on parle d'interaction moléculaire, cela signifie qu'un gène (ou une protéine) interagit avec un autre gène (ou une protéine) ou qu'un gène interagit avec lui-même (on parle alors d'autorégulation). Un filtre permettant de ne retenir que les phrases qui contiennent uniquement deux noms de gènes est en fait la condition minimale requise pour qu'une phrase traite potentiellement d'interaction. Pour effectuer ce filtrage mais aussi pour favoriser les traitements ultérieurs, il est indispensable de mettre en évidence les noms de gènes et de protéines présents dans les différentes phrases. Il suffit pour cela d'apposer devant chaque nom de gène ou de protéine un signe distinctif afin qu'ils puissent être reconnus de façon simple et rapide. Cette étape est nommée balisage.

L'étude de la liste des noms de gènes présents dans la base de données a permis de mettre en évidence que le balisage ne pouvait pas se faire totalement de façon automatique. En effet, certains noms de gènes et de protéines sont ambigus. C'est à dire qu'ils sont soit identiques à des mots vides (exemple : if, is, on, for), soit identiques à des mots du langage courant (exemple : blood, eye, beat). Il a fallu trouver un moyen pour permettre de faire la différence entre les noms de gènes et de protéines et ceux du langage courant. La liste des noms de gènes ambigus a tout d'abord été établie. Pour chacun de ces noms de gènes, une interrogation a été

faite sur la base de données FlyBase. Si celui-ci possède très peu d'information phénotypique c'est qu'il a peu été étudié et donc qu'il y a peu de chance qu'il apparaisse en interaction avec un autre gène. Il n'est donc pas nécessaire de le prendre en compte. De même si ce gène n'est pas présent en tant qu'entrée dans la nouvelle sous-base, il n'est pas nécessaire de le prendre en compte car il sera peu présent. Pour le reste des gènes ambigus chacune des phrases de la sous-base est lue. Si le nom ambigu est un nom de gène, alors un signe distinctif est apposé devant lui.

Synopsis et évaluation

Toutes les étapes de transformation du fichier d'origine, y compris le filtrage des phrases contenant uniquement deux noms de gènes ont permis d'éliminer une bonne partie de l'information non pertinente. La base de données initiale (FlyBase) a une taille de 20 Mo et contient 90000 références. Toutes les étapes de reformatage et de filtrage des données ont permis d'obtenir une nouvelle base de données ne contenant plus que 1200 références et répertorie plus que 550 gènes. Une validation de ces données par un expert du domaine a montré que un peu plus de 50% des phrases du nouveau corpus traitent réellement d'interactions géniques. Cela n'est bien évidemment pas suffisant. Des processus complémentaires doivent donc être appliqués pour arriver à plus de 90% de phrases qui traitent réellement d'interactions et cela sur du texte intégral.

Traitement sémantique des données

Les méthodologies complémentaires visent à étudier le vocabulaire contenu dans chacune des phrases. Celles-ci sont basées sur des techniques d'analyse statistique de données textuelles [Warmesson 1993]. Malgré les outils de pré-traitement qui existent actuellement, les résultats des analyses statistiques des données textuelles sont souvent erronés à cause des problèmes de variété des formes graphiques présentes dans les corpus. Il est donc indispensable, avant d'entreprendre tout traitement statistique sur un corpus de texte, de soumettre les formes graphiques à une lemmatisation, c'est à dire de regrouper dans de mêmes unités, les formes graphiques qui correspondent aux différentes flexions d'un même lemme [Lebart 1994]. A l'aide d'un expert du domaine, cette étape consiste en une homogénéisation des formes.

La lemmatisation du corpus est faite à l'aide d'un dictionnaire comportant plus de 50000 termes du langage courant anglais édité par le producteur de la base de données Derwent. Celui-ci a été complété d'une liste de termes plus spécifiques du domaine biologique. Tous les termes du langage courant qui n'ont pas de signification pour notre analyse, (comme les termes and, for ou the) sont ensuite éliminés. Ces termes sont aussi appelés « mots vides ».

Avec les opérations de lemmatisation, le nombre de formes différentes a été réduit à 1900.

Gene symbol abd-A
Full name abdominal A
FlyBase ID number FBgn0000014
Synonym(s) BXC
Genetic map position 3-58.8
Prosite protein domains
PS00027 == `Homeobox' domain signature.
PS00032 == `Homeobox' antennapedia-type protein signature.
DNA/RNA accessions L31790 X54453
Protein accessions SWP/P29555 PIR/A35915
>>
Data from ref. FBrf0047928
Phenotypic information
The salm gene acts independently of abd-A.

Data from ref. FBrf0049619
Phenotypic information
Expression domains of abd-A have been identified in the midgut visceral mesoderm and the domain position defined with respect to parasegment boundaries.
>>
Data from ref. FBrf0054607
Phenotypic information
trx is necessary for normal levels of abd-A protein accumulation.

Data from ref. FBrf0055051
Phenotypic information
The dosage of abd-A modulates the abdominal leg phenotype in Ubx mutants.
>>
Allele abd-A[39]
Synonym(s) Hab[rev28390.39]
FlyBase ID number FBal0000074
Discoverer(s) R.H. Baker
Mutagen ethyl nitrosourea
>>
References
Bender and Green, 1963, Int. Congr. Genet. 11 1: 173 [FBrf0015474]
Lewis, 1968, Int. Congr. Genet. 12 1: 96--97 [FBrf0019815]
Kiger, 1976, Dev. Biol. 50: 187--200 [FBrf0028540]

données générales

références biblio-graphiques courtes

mutants

liste des références

Figure 1: Exemple d'une entrée de gène dans la base FlyBase

Les noms des champs sont situés à gauche.
Le symbole ">>" signifie que l'entrée n'est pas complète. Certaines informations ont été supprimées.

ANALYSE DES DONNEES

L'objectif de cette étape est de pouvoir reconnaître, avec le plus faible taux d'erreur possible, des phrases décrivant des interactions moléculaires. Pour cela différentes techniques de statistique textuelle sont employées afin de rechercher quels sont les mots ou combinaisons de mots qui sont le plus souvent utilisés dans des phrases décrivant une interaction.
Les techniques utilisées ont été appliquées sur un échantillon de 1199 phrases qui contiennent toutes seulement deux noms de gènes. 653 d'entre elles traitent réellement d'interactions, 491 décrivent autre chose qu'une interaction et 55 d'entre elles sont indéterminées (ne fournissent pas assez d'éléments pour juger sis celles-ci traitent ou non d'interaction.

AFC (Analyse Factorielle des Correspondances)

Une première analyse multidimensionnelle (Analyse Factorielle des correspondances) sur 455 des 1582 formes lemmatisées. (nous n'avons retenus que les formes de fréquence strictement supérieure à 4) a été faite. Le vocabulaire recherché doit permettre d'extraire d'un corpus de documents un maximum de phrases qui traitent potentiellement d'interaction. Il n'est donc pas nécessaire de retenir des termes de faibles fréquences. Pour l'analyse factorielle une matrice contenant des termes de fréquence > 4 (qui sont donc présents dans les phrases analysées au moins 5 fois quelque soit le type de phrase) est donc suffisant.
Cette analyse a permis l'émergence, d'une part, d'une liste de termes spécifiques pour décrire une interaction et de l'autre, une liste de termes relatifs aux phrases qui ne traitent pas d'interaction (fig.2).

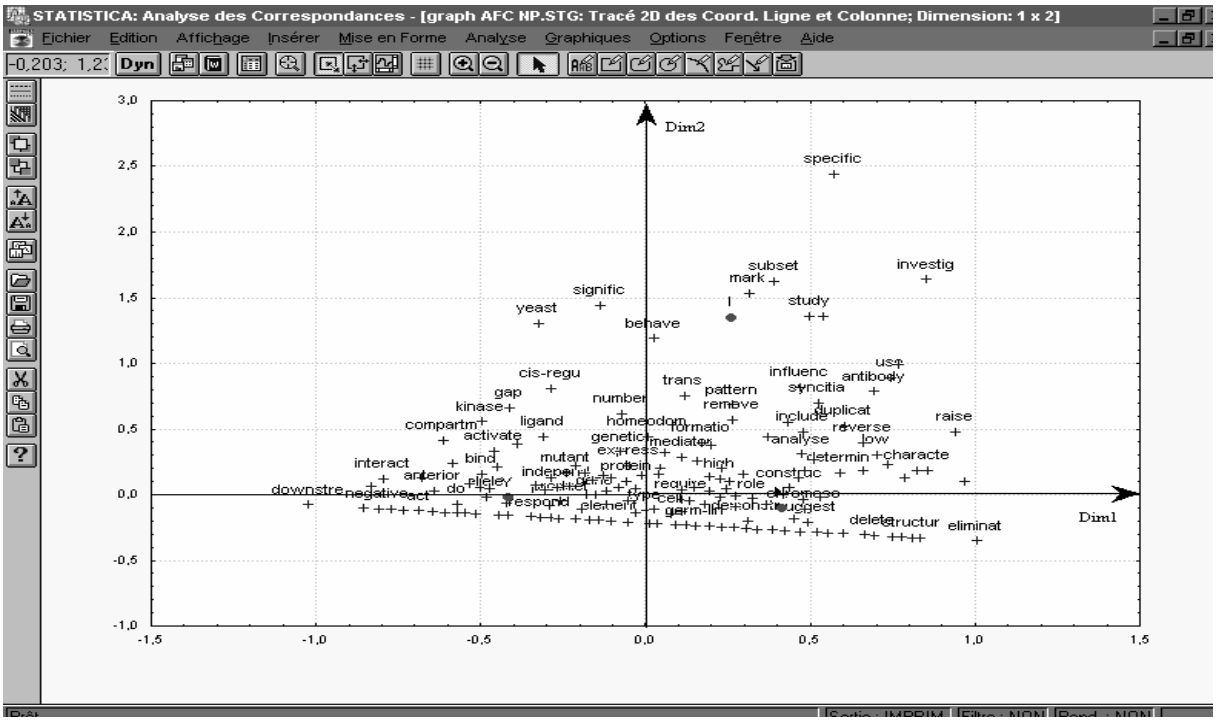


Figure 2: Résultat AFC

Une analyse sémantique (replacer les termes dans leur contexte biologique et dans l'objectif de notre recherche) de chacun de ces termes a été réalisée. Une vingtaine de termes ont été supprimés car ils n'ont pas de sens « réel ». Par contre, autoregulate, derepress, extert, interact

sont des termes qui ont été retenus car ils sont très présents dans les phrases qui traitent d'interactions et presque absent dans les autres types de phrase.

IMFR

En parallèle une deuxième analyse a été faite à l'aide du système « IMFR » développé par un étudiant du CRRM : Bader Roudani. Il a pour objectif d'identifier les multiformes communes répétées [Betaille 1998] dans un corpus textuel et de les ranger dans un fichier structuré que l'on pourra ensuite, suivant les besoins, consulter à l'aide d'un gestionnaire de base de données. Les 1200 phrases traitées par le système à donnée des résultats intéressants. C'est à dire que nous avons pu retenir différents groupes de termes (en majorité des bitermes) qui permettent lors d'une interrogation de la base des 1200 phrases de ne retenir que celles qui traitent réellement d'interaction. Voici quelques exemple de bitermes retenus :
Require – express ; product – require ; transcribing – gene ; protein - function

La superposition de ces deux analyses a permis de mettre en place une série de 39 requêtes de type booléennes (utilisation de ET et OU) retenant des 1199 phrases interrogées, 87% des phrases qui traitent réellement d'interaction avec une pollution de 47 ,5% de phrases qui ne traitent pas d'interaction (fig.3).

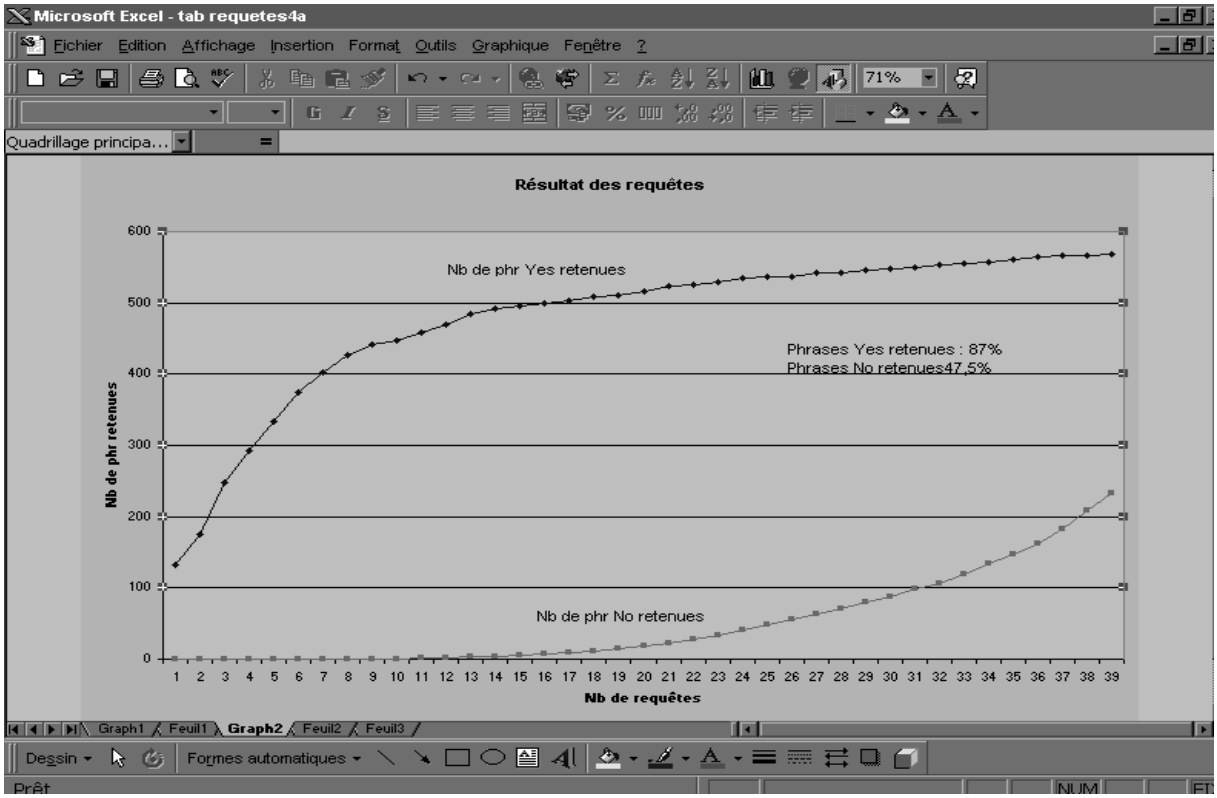


Figure 3 : Résultat des requêtes

Pondération des termes issus de l'AFC

Par ailleurs d'autre études statistiques ont été faites à partir des résultats de l'AFC. Celles-ci consistent à donner un poids pour chacun des termes issus de l'AFC. Un poids positif pour les termes issus de la liste des termes spécifiques pour décrire une interaction et un poids négatif pour le reste des termes. Chaque terme est ensuite remplacé par son poids dans chacune des 1200 phrases (le reste des termes qui n'ont pas été traités par l'AFC sont supprimés). Puis une somme ou une moyenne des poids est effectuée dans chacune des phrases. L'étude de la

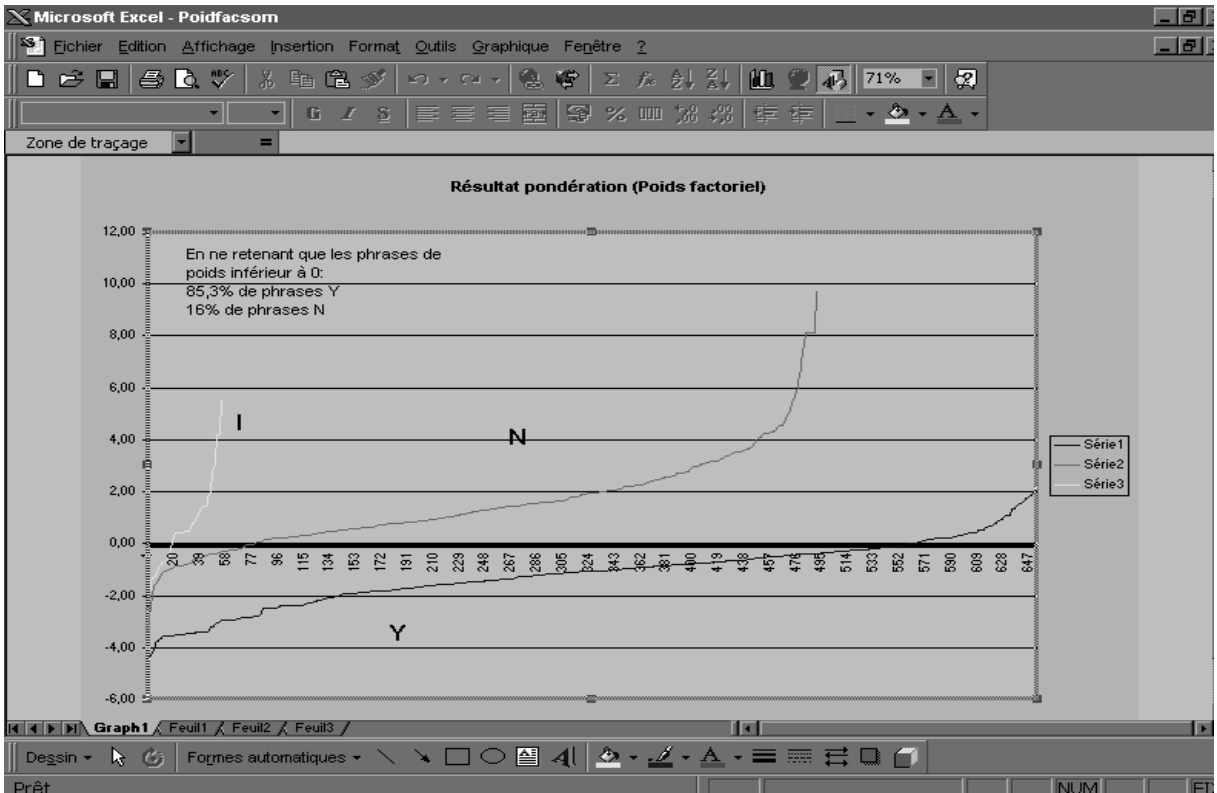
courbe de répartition des poids des phrases en fonction de leur type (phrases qui traitent d'interaction et phrases qui ne traitent pas d'interaction) a permis de déduire de bons résultats. Si l'on ne retient que les phrases qui ont une poids positifs (supérieur ou égal à zéro), nous obtenons 87% des phrases qui traitent réellement d'interaction mais avec une pollution de seulement 19% de phrases qui ne traitent pas d'interactions.

Poids factoriel

Une autre technique de pondération (index de vraisemblance d'interaction) des termes est celle du poids factoriel. Chaque terme issus de l'analyse factorielle possède des coordonnées. Il suffit de projeter chacun de ces termes sur l'axe de dimension 1 (c'est celui qui permet de d'interpréter le pôle Yes et le pôle No). Chaque coordonnées sur cette abscisse correspond alors à un poids : le poids factoriel. Tous les termes issu de l'AFC possèdent alors un poids. Chaque terme est remplacé par son poids dans les 1199 phrases et les termes sans pondération sont supprimés (les termes proches du pôle Yes ont une coordonnée négative et les termes proches du pôle No ont une coordonnée positive). Une somme ou une moyenne des poids est ensuite effectuée pour chacune des phrases. L'étude de la courbe de répartition des poids des phrases en fonction de leur type a permis de déduire de bons résultats. Si l'on ne retient que les phrases qui ont une poids négatif (inférieur ou égal à zéro), nous obtenons 85% des phrases qui traitent réellement d'interaction avec une pollution de 16% de phrases qui ne traitent pas d'interactions (fig.4).

Les différentes techniques de pondération utilisées ci-dessus permettent d'obtenir de bons résultats pour extraire d'un corpus de documents les informations relatives aux interactions. Le but étant de retenir un maximum de données traitant d'interaction avec un minimum de bruit (toute information non relative aux interactions génétiques et moléculaires).
Il s'agira par la suite de combiner ces différentes techniques (allier la statistique et le côté sémantique des termes analysés) pour obtenir les meilleurs résultats possibles.

Figure 4 : Résultat pondération (poids factoriel)



Transposition à d'autres bases de données

Maintenant que la liste de requêtes est en partie créée, il s'agit de vérifier si celle-ci fonctionne aussi bien sur des références sur la Drosophile mais dans la base de données Medline. Pour cela il suffit de retrouver dans Medline les références qui ont été analysées dans FlyBase. La base de données Medline sur le Web donne un lien entre le numéro de référence Medline et le numéro de référence de FlyBase (FBrf). La liste des liens existant entre les 2 bases de données a donc été télédéchargée. Sur les 730 références différentes de FlyBase (issues des 1199 phrases validées), seulement 538 ont un lien vers Medline. Il suffit de créer un nouveau fichier qui comprend pour chaque numéro de référence Medline, le résumé Medline et les phrases de FlyBase correspondant à ce résumé plus le champ validation de chacune de ces phrases.

REPRESENTATION GRAPHIQUE DES DONNEES

Acquisition d'une liste d'interactions

Un logiciel de traitement infographique (matrisme) [Matrisme] [Boutin 1996] permet de visualiser les réseaux des interactions des bases de données nouvellement créées sous forme de graphe. Le résultat de requêtes simples (quels sont les gènes contrôlés par le gène X, quels sont les gènes régulant les gènes Y, …) pourront être obtenues et les informations complémentaires associées aux gènes et interactions seront ensuite accessibles, à partir du graphe, grâce à des liens hypertexte. Toutes ces informations seront stockées dans une base de données : KNIFE [Tropserver 1998] (Knowledge on Networks of Interactions in the Fly and other Eukaryotes) dont une partie est déjà accessible sur Internet. Ce site contient actuellement des informations de base sur une centaine de gènes en interactions avec des liens vers d'autres bases de données telles que FlyBase, Genbank, Swiss-Prot ou Medline. La page d'accueil de cette base est représentée en figure 5.

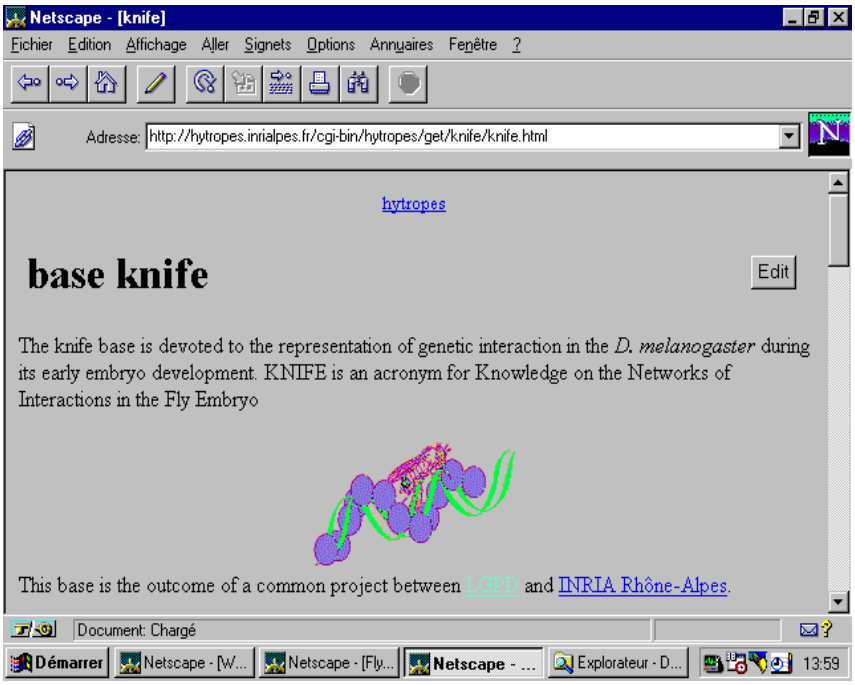


Figure 5 : page d'accueil de la base Knife

Un essai de représentation graphique de ces réseaux a déjà été initié. Les 1200 phrases extraites de la base de données FlyBase ont été validées et 650 d'entre elles traitent réellement d'interactions. Une extraction des noms de gènes présents dans ces phrases a permis de déterminer les relations existantes entre ces différents gènes. Les 650 phrases positives répertorient 334 gènes différents créant ainsi 452 interactions différentes et réelles. Celles-ci forment 46 réseaux différents et indépendants les uns des autres. Chaque réseau est constitué en moyenne de 3 gènes différents sauf un qui regroupe 215 gènes différents. En plus de ces 46 réseaux il existe 4 autorégulation :

rl → rl

r → r

puc → puc

Ace → Ace

Plus de 45 réseaux d'interactions différents incluant 334 gènes différents ont été déduits. La figure 6 représente l'image d'un de ces réseaux obtenus à partir de ces phrases.

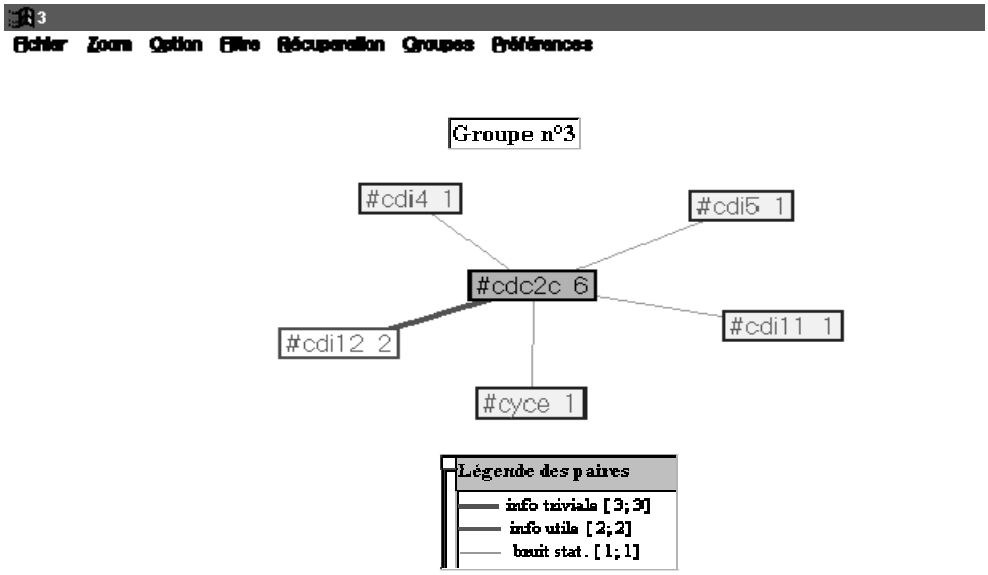


Figure 6 : Image d'un réseau d'interactions chez des gènes de la drosophile

CONCLUSION ET PERSPECTIVES

Les règles de reconnaissances établies pour extraire les données sur les interactions génétiques et moléculaires seront ensuite testées et affinées sur une nouvelle source de données : Medline. Cette source d'informations présente l'avantage d'être d'une part très complète (plus de 8 millions de résumés d'articles) et d'autre part, de stocker des informations concernant plusieurs organismes. Nous pourrons élargir notre analyse sur des données de la drosophile mais aussi sur la souris et une partie du génome humain. Toutes les données extraites permettront d'alimenter de nouvelles bases de données structurées, dédiées aux interactions. Il sera alors possible, à partir du logiciel infographique Matrisme, d'établir une représentation graphique du réseau que forment ces interactions. L'analyse de chacun des réseaux permettra de définir leurs caractéristiques et de voir s'il existe des régularités ou des similarités entre sous-réseaux au sein d'un même organisme. La comparaison évolutive entre graphes d'interactions de l'homme et d'organismes modèles permettra de mieux étudier sur l'animal des interactions qui ne peuvent l'être chez l'homme et ainsi de compléter nos connaissances sur les réseaux d'interactions moléculaires de l'homme.

Ces bases de connaissances et ces graphes de réseaux établis de façon semi-automatique seront à la fois utiles pour les laboratoires de recherche (laboratoire de génétique...) et pour les industriels (entreprise pharmaceutique, de biotechnologie…). Cette méthodologie devrait faire apparaître et valoriser de nouvelles informations afin de mieux cibler et optimiser la recherche et aussi de fournir aux laboratoires des pistes de recherche complémentaires.

BIBLIOGRAPHIE

[Betaille 1998] Betaille H., Massotte A.M., Joubert A., Recherche de similitudes entre fragments de documents, JADT, Nice 1998.

[Boutin] Boutin E., Dumas S., Rostaing H., Quoniam L., Les réseaux comme outil d'analyse en bibliométrie. Un cas d'application : les réseaux d'auteurs. Les cahiers de la documentation belge, 1, 3-13, 1996.

[FlyBase 1996] US National Institute of Health and the British Medical Research (1996), *FlyBase @ FlyBase.bio.indiana.edu*. [Online]. URL adress : http://flybase.bio.indiana.edu/

[Infotrans] Infotrans : logiciel de reformatage et de dédoublonnage développé par la société I+K

[Lebart 1994] Lebart et Salem, Statistique textuelle, ed. Dunod, 1994.

[Lindsley 1992] Lindsley D.L. and Zimm G.G., The Genome of Drosophila melanogaster, p. 1133, New York, Academic Press, 1992.

[Matrisme] Matrisme : logiciel de traitement infographique développé par Eric Boutin dans le cadre de sa thèse au CRRM et L.E.P.O.N.T

[Tropserver 1998] Tropeserver (1998), *Knife*. [Online]. http://hytropes.inrialpes.fr/cgi-bin/hytropes/get/knife/knife.html

[Warmesson 1993] Warmesson I., Parisot P., Bedecarrax C., Huot C., Traitements linguistiques et analyse des données pour une exploitation systématique des banques de données, Revue Française de Bibliométrie Appliquée, 12, 281-294, 1993.