Sciences de l'information pour l'étude des systèmes biologiques (exemple du vieillissement du système immunitaire)

Walid Bedhiafi

Université Pierre et Marie Curie

Université de Tunis El-Manar

Ecole Doctorale Complexité du vivant - ED515

Laboratoire Immunologie, Immunopathologie, et Immunothérapie - UMRS 959

Laboratoire de Génétique, Immunologie et Pathologies Humaines – LR05ES05

Sciences de l'information pour l'étude des systèmes biologiques

(Exemple : vieillissement du système immunitaire)

Par Walid BEDHIAFI

Thèse de doctorat de bioinformatique

Présentée et soutenue publiquement le 20 Septembre 2017

Devant un jury composé de :

Dr. Frédérique PERONNET	Présidente
Pr. Ahmed REBAI	Rapporteur
Dr. Amel BORGI	Rapporteure
Dr. Olivier DAMERON	Examinateur
Pr. Adrien SIX	Directeur
Pr. Amel BENAMMAR EL GAAIED	Directrice
Dr. Véronique THOMAS-VASLIN	Encadrante

Programme Doctoral International Modélisation des Systèmes Complexes

PDI MSC (IRD-UPMC)

« Les explications réductionnistes ont atteint leurs limites, pour progresser, il faudrait de nouvelles approches pour traiter les données existantes, au lieu d'accumuler encore davantage de données. »

Stephen Jay Gould

A la mémoire de celui qui fut mon maître d'école, Amor CHERICHI…

A mes parents et ma sœur dont la vie n'a pas été toujours facile…

Remerciements

Mon jury de thèse

Docteur Frédérique PERONNET, je tiens à vous remercier d'avoir accepté de présider ce jury et d'avoir consacré du temps pour discuter et évaluer ce travail.

Professeur Ahmed REBAI, je tiens à vous remercier d'avoir accepté de juger de la qualité de ce travail**.** Vos remarques et suggestions ont contribué objectivement à son amélioration.

Docteur Amel BORGI, je vous remercie d'avoir accepté de faire partie de ce jury et d'avoir donné de votre temps pour lire ce manuscrit, malgré vos nombreuses charges. Vos remarques et vos suggestions m'ont été très utiles.

Docteur Olivier DAMERON, je vous remercie pour tout ce que vous avez fait pour moi depuis le master et pour le temps que vous avez pris pour évaluer ce travail. Vos conseils avant et tout au long de cette thèse m'ont été précieux. Je ne pourrais oublier que c'est grâce à vous que j'ai été initié aux sciences de l'information et surtout aux ontologies biologiques.

Mes directeurs de thèse

Professeur Adrien SIX, je vous suis reconnaissant pour m'avoir ouvert les portes de votre équipe au sein du laboratoire i3. J'ai beaucoup appris à vos côtés. Je tiens à vous remercier pour vos conseils qui m'ont orienté tout au long de ces années, pour votre patience, pour votre disponibilité malgré la distance et pour tout le temps que vous m'avez consacré.

Professeur Amel BENAMMAR ELGAAIED, il est très rare de rencontrer une personne d'une telle ouverture d'esprit. J'en profite pour remercier Néjla qui a permis cette rencontre. Je vous serai toujours très reconnaissant, pour votre patience, vos encouragements et surtout pour vos conseils. Je vous serai toujours très reconnaissant de m'avoir ouvert les portes de votre laboratoire. Je vous remercie pour ces quatre années passées à vos côtés, pour les collaborations que vous m'avez permis. Votre départ à la retraite sera une perte pour nous tous.

Docteur Véronique THOMAS-VASLIN, je vous remercie d'avoir accepté de co-encadrer ce travail. Vos remarques ont toujours été justes, elles m'ont beaucoup aidé à avancer. Merci d'avoir permis toutes les collaborations et de m'avoir ouvert l'esprit sur d'autres disciplines. Merci aussi de votre disponibilité et du temps que vous m'avez consacré malgré la distance.

Mon tuteur de thèse

Professeur Hugues BERSINI, je vous remercie de votre disponibilité malgré la distance, de vos conseils et remarques qui m'ont orienté tout au long de ce projet.

Le laboratoire i3

Professeur David KLATZMANN, je vous remercie de m'avoir accepté au sein de vos équipes. J'ai beaucoup appris en évoluant dans votre laboratoire.

Docteur Wahiba CHAARA et Docteur Nicolas DERIAN, enfin, Wahiba et Nico☺, mes collègues de bureau, mes conseillers en cas de crise et surtout mes correcteurs de français : merci pour vos conseils, pour vos encouragements, et merci surtout d'avoir facilité mon intégration parmi vous.

Docteur Hang-Phuong Pham, merci pour ta disponibilité pour tous tes conseils et pour ton soutien.

Docteur Encarnita FERRANDIZ-MARIOTTI, Docteur Iannis DRAKOS, Docteur Valentin QUINIOU et Docteur Claude BERNARD, merci de m'avoir accepté parmi vous et pour toutes les conversations qu'on a pu avoir sur la science ou sur la vie autour des friandises que vous apportez ou autour du canard enchaîné de Claude. Ces quatre ans sont vite passés.

Sophie Miller et Sandrine THOMAS, merci pour le support que vous m'avez apporté, votre aide m'a été d'un grand secours.

Et pour tous les collègues du laboratoire que je m'excuse de ne pas citer, je vous expime tout mes remerciement pour le plaisir que vous m'avez offert de travailler avec vous.

Le laboratoire LGIPH

Docteur Nejla STAMBOULI, merci de m'avoir introduit au laboratoire, merci pour ta présence et pour ton soutien infaillible…

Sami, Farouk et Chéma, je vous remercie pour votre support, pour tous les fous rires qu'on a pu avoir, pour votre aide pour tous les projets qu'on a entrepris ensemble. Votre soutien m'a été d'un grand secours.

Docteur Ines OMRANE, tu as toujours été là quand j'ai eu besoin d'un coup de main, merci de ta disponibilité de ton soutien, bonne chance et bonne continuation là où tu es.

Khouloud HAMDI, tu as été toujours à l'écoute, merci pour tout.

Docteur Lotfi CHERNI, merci pour tous tes conseils et ton soutien. J'ai beaucoup apprécié nos discussions

Mes collègues et mes professeurs du conseil du laboratoire, je vous remercie pour votre soutien, et d'avoir été là en cas de besoin.

Fadoua, Héla, Monia, Rym, Houcine, Sonia, Senda, Yassine, et tous mes autres collègues du laboratoire que je m'excuse de ne pas citer, je vous exprime tout mes remerciement pour le plaisir que vous m'avez offert de travailler avec vous, pour avoir facilité mon intégration parmi vous, pour votre soutien et pour tous les fous rires qu'on a pu avoir. Quatre années sont passées vite.

Le Laboratoire LIPAH

Professeur Sadok BEN YAHIA, je vous remercie pour tout le temps que vous m'avez consacré et pour tout l'aide que vous m'avez apportée.

Marwa BEN MBAREK, merci d'avoir accepté la collaboration, merci de ta disponibilité, j'espère qu'on va continuer.

Le programme Doctoral International

Docteur Christophe CAMBIER et Docteur Jean-Daniel ZUKER, je vous remercie pour cette opportunité et pour tous les efforts que vous avez fournis pour la réussite de ce programme.

Mes amis du PDI, Docteur Sliman BEN MILED, Dorra, Ghassen, Karim, Hédia, Diarra, Malek et tous les autres, merci pour tous les échanges qu'on a pu avoir.

Mes amis, et vous êtes nombreux, les deux **Amine, Wided, Wanessa, Hazem, Ghazi**, **Belhassen** et tous ceux que j'ai oublié de citer je vous remercie de votre soutien indéfectible tout aux longs de ces années.

Résumé

Les approches expérimentales à haut débit pour l'étude du transcriptome impliquent plusieurs étapes de traitement pour la quantification de l'expression des gènes et pour l'annotation qui permet l'interprétation des résultats. Le laboratoire i3 U 959 (INSERM-UPMC) au sein duquel ce projet a été effectué en cotutelle avec le laboratoire LGIPH LR05ES05 (UTM), utilise ces approches dans ces différentes thématiques. Des limites ont été observées quant à l'utilisation des approches classiques pour l'annotation des signatures d'expression des gènes. L'objectif principal de cette thèse a été de développer une nouvelle approche d'annotation pour répondre à ce besoin. L'approche que nous avons développée est une approche basée sur la **contextualisation** des gènes et de leurs produits puis sur la **modélisation des voies biologiques** pour la production de bases de connaissances pour l'étude de l'expression des gènes. Nous définissons ici un contexte d'expression des gènes comme suit :

$$\boldsymbol{Contexte = population\ cellulaire + compartiment\ anatomique + état\ pathologique}$$

Pour connaitre les contextes d'expression des gènes, nous avons opté pour le criblage massif de la littérature et nous avons développé un package Python, **OntoContext** (https://github.com/walidbedhiafi/OntoContext1) qui permet d'annoter les textes automatiquement en fonction de trois ontologies choisies en fonction de notre définition du contexte. Nous montrons ici que notre package assure des performances d'annotation textuelle meilleures que **NCBO annotator**, l'outil de référence. Nous avons utilisé OntoContext pour le criblage d'un corpus sur le vieillissement du système immunitaire dont on présente ici les résultats.

Pour la modélisation des voies biologiques nous avons développé en collaboration avec le laboratoire LIPAH une méthode de modélisation basée sur un algorithme génétique qui permet de combiner les résultats de mesure de la proximité sémantique sur la base des annotations des gènes dans l'ontologie *Biological* ***Process*** **et les** données d'interactions de la base de données db-string. Nous avons réussis retrouver des réseaux de références avec un taux d'erreur de 0,47.

Mot clés : expression génétique, information, contexte, ontologie, annotation automatique, fouille de textes.

Abstract

High-throughput experimental approaches for gene expression study involve several processing steps for the quantification, the annotation and interpretation of the results. The i3 U 959 (INSERM-UPMC) lab, in which this project was carried out in co-supervision by the LGIPH LR05ES05 (UTM) lab, applies these approaches in various experimental setups. However, limitations have been observed when using conventional approaches for annotating gene expression signatures. The main objective of this thesis was to develop an alternative annotation approach to overcome this problem.

The approach we have developed is based on the **contextualization** of genes and their products, and then **biological pathways modeling** to produce a knowledge base for the study of gene expression. We define a gene expression context as follows:

$$\boldsymbol{Context = cell\ population + \ anatomical\ compartment + \ pathological\ condition}$$

For the production of gene contexts, we have opted for the massive screening of literature. We have developed a Python package, **OntoContext**, (https://github.com/walidbedhiafi/OntoContext1). OntoContext allows annotating the texts according to three ontologies chosen according to our definition of the context. We show here that OntoContext ensures better performance for text annotation that **NCBO annotator**, the reference tool. We used OntoContext to screen an aging immune system text corpus. The results are presented here.

To model the biological pathways we have developed, in collaboration with the LIPAH lab a modeling method based on a genetic algorithm that allows combining the results semantics proximity using the Biological Process ontology and the interactions data from db-string. We were able to find networks with an error rate of 0.47.

Key words: gene expression, information, context, ontology, automatic annotation, text mining

Liste des abréviations

ACP	Analyse en Composante Principale	
ADN	Acide Désoxyribonucléique	
API	Interface de programmation	Application Programming Interface
ARN	Acide Ribonucléique	
ASIS&T	Société américaine des sciences de programmation et de technologie	American Society of Information Science and Technology
BP	Processus biologique	Biological Process
CNN	Réseau neuronal convolutif	Convolutional Neural Network
DAG	Graphe acyclique orienté	Direct Acyclic Graph
Datamining	Fouille des données	
GEO	Omnibus d'expression des gènes	Gene Expression Omnibus
GO	Ontologie des gènes	Gene Ontology
GOA	Annotation de l'ontologie des gènes	Gene Ontology Annotation
HCL	Regroupement hiérarchique	Hierarchical clustering
HGNC	Comité de nomenclature des gènes	Human Gene Nomenclature Committee
HTML	Langage de balisage hypertexte	HyperText Markup Language
http	Protocole de transfert hypertexte	HyperText Transfert Protocol
I3	Laboratoire d'Immunologie, Immunopathologies, Immunothérapie	
IC	Contenu en information	Information Content
IPA		Ingenuity Pathway Analysis
KDD	Découverte de connaissance à partir des bases de données	Knowledge Discovery in Databases
KEGG	Encyclopédie de Kyoto des gènes et des génomes	Kyoto Encyclopedia of Genes and Genomes
LCS	Ancêtre commun le plus bas	Lowest common ancestor
LGIPH	Laboratoire de Génétique, Immunologie et Pathologies Humaines LGIPH	
LINCS	Librairie des réseaux intégrés de gènes basés sur les signatures cellulaires	Library of Integrated Network-based Cellular Signatures
LIPAH	Laboratoire en Informatique en Programmation Algorithmique et Heuristique	

MeSH	Rubriques médicales	Medical Subject Headings
MICA	Ancêtre commun le plus informatif	Most Informative Common Ancestor
NCBI	Centre National de l'information biotechnologique	National Center of Biotechnology Information
NLT	Boite à outils du langage naturel	Natural Language Toolkit
OWL	Langage Web des Ontologies	Ontology Web Language
POS Tagging		Part Of Speech Tagging
RDF	Système de description des ressources	Resource Description Framework
SVM	Machine à vecteurs de support	Support Vector Machine
TiGER	L'expression et la régulation des gènes dans les tissus	Tissue Gene Expression and Regulation
UMLS	Système de langage médical unifié	Unified Medical Language System

Table des matières

Avant-propos

Dans cette thèse nous nous intéressons à l'extraction à partir de la bibliographie de données relatives à l'expression des gènes -dans un contexte biologique donné- : l'emplacement dans le corps (localisation anatomique), le type cellulaire (population cellulaire) et la maladie associée (état pathologique) tout en donnant du sens aux données extraites en termes de processus et voies biologiques. Pour nous assurer de la bonne lecture du document il est indispensable de définir quelques termes au début de ce manuscrit. Les définitions seront par la suite abordées avec plus de détail au fûr et mesure de l'avancement dans le manuscrit.

- **L'information** : est un ensemble de données ayant un sens. Le sens est défini par le contexte. Dans cette thèse l'information étudiée concerne l'expression des gènes.
- **Le contexte** : en science de l'information, il constitue l'ensemble des circonstances qui accompagnent une information et qui lui donne du sens. Dans ce cas précis il s'agit du contexte cellulaire, anatomique et pathologique de l'expression des gènes.
- **La contextualisation** : est le processus qui permet d'associer une donnée à un contexte. Dans notre cas il s'agit d'associer un gène ou plutôt une liste de gènes à un compartiment anatomique, une population cellulaire et un état pathologique.
- **L'annotation des expériences de transcriptomique** : est le processus qui permet d'associer un gène ou une protéine à un processus biologique ou une pathologie par le criblage des bases de données spécialisées.
- **L'annotation textuelle** : est le processus qui permet d'identifier soit des mots clés soit des expressions d'intérêt dans un texte ou un corpus de textes. Dans cette thèse nous avons utilisé les résumés d'articles scientifiques issus d'une requête PubMED. Nous avons utilisé, comme référence pour cette tâche, des dictionnaires issus des ontologies biologiques.
- **Une ontologie** : est un vocabulaire contrôlé et structuré de manière hiérarchique pour décrire un domaine de connaissance donné. Dans cette thèse nous avons utilisé une ontologie pour les populations cellulaires, une ontologie pour les localisations anatomiques, une ontologie des pathologies humaines et une ontologie des processus biologiques.
- **Une voie biologique** : est un ensemble de gènes qui interagissent ensemble pour assurer une fonction biologique ou un processus cellulaire.
- **Un corpus de textes** : est un ensemble de textes relatif à un domaine d'étude particulier. Dans cette thèse nous avons deux types de corpus : les corpus de validation qui sont des

textes pré annotés par des experts et les corpus d'étude que nous avons générés par requête PubMED.

- **Une requête PubMED** : est une requête lancée dans la base de données bibliographique MEDLINE par des mots clés dont le résultat est une liste de résumés d'articles scientifiques
- **Les MeSh Terms** : constituent un vocabulaire contrôlé utilisé pour annoter les résumés d'articles dans MEDLINE. Ce sont des mots clés qui peuvent être utilisés pour la recherche bibliographique par requêtes PubMED.
- **Fouille des données (Data mining)** : est un ensemble d'algorithmes basée sur des modèles inspirées de disciplines diverses tel que les statistiques et l'intelligence artificielle pour extraire de la connaissance à partir de volumes importants de données.

Introduction

A. Introduction générale

Les deux laboratoires d'accueils (Laboratoire de Génétique, Immunologie et Pathologies Humaines à Tunis et le laboratoire Immunologie, Immunopathologies, Immunothérapies à Paris) s'intéressent à l'étude du système immunitaire, un système biologique complexe. Un axe de recherche du laboratoire Immunologie, Immunopathologies, Immunothérapie concerne le vieillissement du système immunitaire. Dans le cadre de ces recherches, des données multiparamétriques ont été produites par des technologies à haut débit. Un objectif de cette thèse est le développement d'une approche d'annotation (interprétation) de ces données. Pour ce faire nous allons utiliser des techniques des sciences de l'information et des données.

Le premier chapitre est un chapitre introductif dans lequel nous allons définir et exposer l'évolution de l'information et des technologies de l'information et les approches utilisées pour le traitement et l'analyse de l'information pour la production de nouvelles connaissances. Je détaille aussi quelques applications. Les systèmes biologiques sont des systèmes informatifs et générateurs d'informations (comme on va l'expliquer dans ce chapitre), renforcés par l'évolution des technologies expérimentales. En effet, l'essor des technologies omiques (séquençage à haut débit et micropuces) a induit une augmentation exponentielle de la production scientifique et des données disponibles. Cette évolution technologique et informationnelle a modifié les pratiques d'analyse des données. Néanmoins, le challenge qui nous est posé concerne l'interprétation des résultats expérimentaux massifs. L'automatisation du processus d'analyse est une réponse ; ainsi, les technologies de l'information et des données peuvent nous aider dans ces approches. Je présente ici deux applications qui ont été développées dans ce sens.

Dans le second chapitre, je présente la première étape d'une démarche globale que nous avons développée pour l'étude des résultats d'analyse de données de transcriptome. Notre démarche est basée sur le criblage de la littérature pour la production d'une liste de signatures d'expression de gènes. La signature est une liste de gènes/protéines contextualisée. Le contexte est défini par une population cellulaire, un compartiment anatomique (un tissu ou un organe…) et/ou une pathologie. Pour ce faire, nous avons développé un outil d'annotation textuelle, OntoContext sous la forme d'un package Python. Dans ce chapitre également nous présentons les étapes de développement et les résultats de validation de cet outil. OntoContext utilise trois ontologies biomédicales pour l'annotation du contexte, et *GENIA tagger* un package Python

disponible basé sur un modèle d'apprentissage pour l'annotation des noms de gènes. Nous présentons enfin le résultat d'annotation d'un corpus, centré sur le vieillissement du système immunitaire, de plus de 120 000 résumés d'articles, extrait de MEDLINE (la base de données bibliographiques).

Le troisième chapitre présente, une approche de modélisation des réseaux biologiques qui a été développée avec le Laboratoire en Informatique en Programmation Algorithmique et Heuristique. A partir de la base de signatures présentées dans le second chapitre, nous cherchons à modéliser des réseaux biologiques (ensemble de gènes qui interagissent pour mener une fonction biologique). Pour ce faire, nous utilisons une base de données d'interactions et les données de similarité sémantiques calculées à partir de l'ontologie *Biological Process* de *Gene Ontology*. Le laboratoire d'informatique a produit un algorithme génétique que nous avons validé. Nous avons opté pour cette méthode suite à des expériences que nous avons menées pour l'utilisation des distances sémantiques pour la modélisation des réseaux de gènes.

La dernière partie est consacrée à la discussion des résultats obtenus pour démontrer l'intérêt de la démarche proposée.

B. Délimitation du sujet

Comme mentionné précédemment, les deux laboratoires d'accueil au sein desquels cette thèse a été menée s'intéressent à l'étude du système immunitaire et utilisent des approches basées sur l'analyse du transcriptome (ARNm) et l'analyse des populations cellulaires. Un des axes développés par le laboratoire i3 est l'étude du vieillissement du système immunitaire. Pour ce faire des données multiparamétriques ont été produites : des données cellulaires (comptage de populations cellulaires) et des données transcriptomiques à haut débit (micropuces à ARN) dans le sang et différents organes immunitaires (rate, thymus et ganglions lymphatiques) sur trois lignées de souris, du jeune âge à la sénescence. L'analyse statistique de ces données lors d'un projet doctoral précédent (Pham 2013) a permis l'identification d'une série de marqueurs biologiques liés au vieillissement. Les résultats obtenus lors du précédent projet, montrent que, le vieillissement du système immunitaire se traduit sur un plan cellulaire par une diminution des lymphocytes T naïfs[1], et une perte de la diversité du répertoire TCR[2] et sur un plan moléculaire par une baisse de l'expression des gènes impliqués dans la prolifération. Cette analyse basée sur l'utilisation des sources classiques pour l'interprétation des résultats a montré des limites. La finalité du projet sur le vieillissement du système immunitaire est de produire un modèle multi-échelle (molécule, voies biologiques intracellulaires, cellule, population cellulaire, organe et organisme) qui tienne compte de la complexité du système immunitaire : un système dynamique (diversité des voies de signalisation et différentiation, des voies pour la régulation génétique, des voies métaboliques), un système fluide (communication entre les différents organes, tissus et compartiments cellulaires). Or, pour la production de ce modèle, on a besoin d'une description fine des phénomènes biologiques qui y participent. D'autres projets sont menés par les deux laboratoires qui utilisent les mêmes approches expérimentales basées sur l'étude des populations cellulaires et l'analyse du transcriptome notamment pour l'étude des maladies auto-immunes[3,4].

[1] Les lymphocytes T sont responsables de l'immunité cellulaire. Les lymphocytes T naïfs sont des lymphocytes qui n'ont pas été activés par les cellules présentatrices d'anagènes.

[2] C'est un complexe moléculaire se trouvant à la surface des lymphocytes T ; il assure une fonction de reconnaissance de l'antigène.

[3] http://www.aidaproject.net/

[4] https://www.transimmunom.fr

Mon projet de thèse s'inscrit dans cette perspective. En effet, notre objectif est d'améliorer les approches d'interprétation des observations expérimentales, pour réaliser des annotations, c'est-à-dire définir les mécanismes mis en jeux dans un contexte biologique donné. Les résultats expérimentaux étant obtenus par des approches à haut débit, ils ont été traités par des méthodes statistiques pour la quantification d'expression des gènes. Or, il nous-est apparu impératif de pouvoir automatiser le processus d'annotation relatif aux aspects qualitatifs et plus précisément au contexte d'expression de ces gènes. Les approches classiques pour l'annotation des expériences sont basées sur l'interrogation des bases de données des voies biologiques et des ontologies. La plupart de ces bases de données telles que KEGG Pathway et Reactôme n'abordent qu'implicitement la notion de contexte biologique ; alors que les ontologies biologiques sont spécifiques chacune à un domaine de connaissance de la biologie sans ou avec peu de liens entre elles. Un besoin de mettre en relations les différentes ontologies permettrait de mieux définir le contexte biologique pour l'expression des gènes. D'autres approches proposent l'utilisation des ressources bibliographiques (voir chapitres 2 et 3) pour l'extraction de connaissances indispensables à l'annotation des données biologiques. Dans ce projet nous abordons l'exemple du vieillissement du système immunitaire qui est un système qui se caractérise par sa complexité et sa nature dynamique et sa nature multi-échelle (de l'échelle moléculaire à l'échelle de l'organisme en passant par les cellules et les organes). Il nous est apparu nécessaire d'envisager une approche méthodologique pour la modélisation automatique des voies biologiques en général tenant compte de l'expression de gènes dans des contextes biologiques précis. A cet effet nous avons développé une approche basée sur les sciences de l'information et les sciences des données pour la contextualisation des gènes et leurs produits (Figure 1) et pour la modélisation automatique des voies biologiques. L'approche proposée ici est une approche à deux étapes :

- Pour la contextualisation, nous procédons à une analyse automatique de la littérature biologique. Pour nous, un contexte est défini par une population cellulaire, une localisation anatomique, et une pathologie. En effet, une des problématiques sous-jacentes du projet est de relier les données transcriptomiques aux données cellulaires et de voir les liens fonctionnels et physiologiques qui pourraient exister entre le vieillissement du système immunitaire et certaines pathologies. **Par l'annotation automatique de la littérature biomédicale**[5], nous cherchons des gènes - et leurs produits - qui ont été co-cités avec des

[5] A ne pas confondre avec l'annotation des expériences. L'annotation textuelle ici est un étiquetage automatique des textes pour l'identification de concepts d'intérêts.

concepts cellulaires, pathologiques et anatomiques. Les problématiques auxquelles nous avons fait face sont liées aux corpus de textes et au choix des terminologies biomédicales à utiliser pour l'annotation (voir chapitre 2). A cet effet nous nous sommes proposés de développer un package Python qui permet d'annoter les textes biomédicaux et de le valider en confrontant ses performances à d'autres outils d'annotation textuelle comme NCBO Annotator en nous basant sur les mêmes corpus de textes. Nous avons appliqué cet outil à un corpus de textes traitant du vieillissement du système immunitaire sélectionné à partir de la base de données PubMED en utilisant une requête avec les MeSH terms spécifiques. Notre objectif final est d'attribuer ces gènes contextualisés à des voies biologiques. Cette étape est toujours en cours de validation avant d'être généralisée.

- A cet effet, nous avons entrepris dans une deuxième partie de cette thèse, une approche de regroupement de gènes qui interviennent dans une même voie biologique. Pour cela nous nous sommes basés d'une part sur les similarités sémantiques établies à partir des termes *Gene Ontology* de l'ontologie Biological Process et d'autre part sur des données d'interactions définies dans la base d'interaction db-STRING. Pour la modélisation de ces réseaux de gènes, nous avons testé deux approches : une approche de calcul de la similarité sémantique couplée à une méthode clustering puis une approche d'apprentissage automatique, sur la base d'un algorithme génétique qui permet de combiner la proximité sémantique et les données d'interaction pour la modélisation des voies biologiques. L'objectif visé est de pouvoir regrouper les gènes qui interviennent dans les mêmes voies biologiques. La question de la modélisation des réseaux de gènes est une autre problématique sous-jacente du projet pour le développement d'une base de connaissances pour la description les mécanismes en jeu au cours du vieillissement (ou un autre phénomène biologique). Or, pour pouvoir atteindre ces objectifs, il faut passer par la description fine des processus moléculaires clés qui interviennent dans les populations cellulaires étudiées (voir chapitre 3).

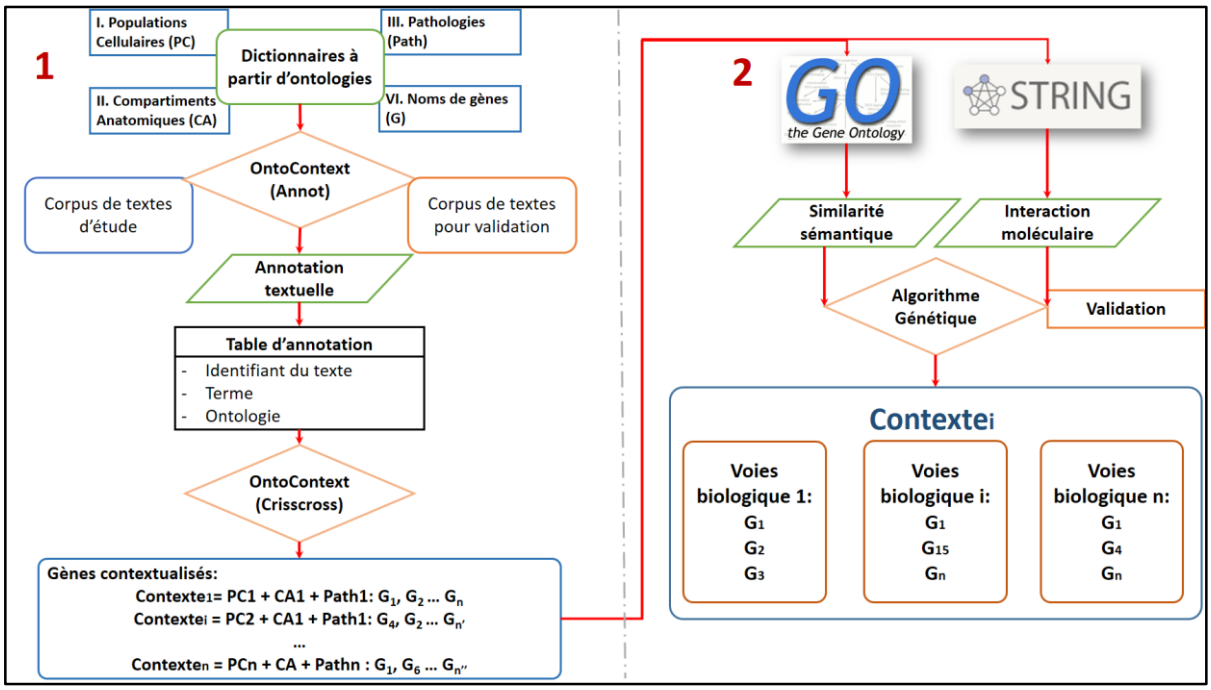


Figure 1. Schéma récapitulatif de l'approche proposée. Il s'agit d'une approche à deux étapes : 1) Une annotation automatique de la littérature pour l'identification des populations cellulaires des compartiments anatomique et des pathologies dans les textes pour la contextualisation des gènes et de leurs produits. 2) Modélisation des réseaux de gènes à partir des listes de gènes contextualisées.

Chapitre 1 : Sciences de l'Information et Biologie (Etat de l'Art)

A. Sciences de l'information

1. Définir l'information

La revue de la littérature montre que la définition du concept « information » est problématique en raison du nombre de disciplines qui l'étudient (des sciences sociales aux sciences exactes telles que les mathématiques). Sur un plan étymologique, le terme information provient du verbe latin *informare* : action de former, de façonner (Rafael and Hjørland 2003, Leleu-Merviel and Useille 2008). A partir de cette recherche étymologique Rafael et Hjorlaand détachent deux définitions liées : l'information comme acte de donner une forme à l'esprit, soit comme acte de communiquer des connaissances. En philosophie, le concept d'information a été associé aux processus de connaissance et d'apprentissage. Ainsi, Descartes définit l'information comme un ensemble d'idées qui donnent forme à une pensée (Leleu-Merviel and Useille 2008). L'avènement de la théorie de l'information par Shannon (Shannon 1948) a été à l'origine d'une nouvelle définition de l'information. En effet, cette théorie s'intéresse à la quantification de l'information. L'information ici devient une entité physique mesurable. C'est une entité transmise entre un émetteur et un récepteur. Dans sa quantification du signal, Shannon s'est intéressé aux « *symboles que porte l'information* », définissant ainsi l'information comme l'ensemble de patterns transmis, et faisant abstraction du contenu sémantique, qu'il « *ne juge pas pertinent* » (Floridi 2005, Leleu-Merviel and Useille 2008). Cette théorie va être à l'origine d'une controverse entre généticiens qui ont introduit la notion de l'information génétique et les spécialistes de la théorie de l'information (Longo et al. 2012). En effet, depuis les expériences de Jacob et Monod (Jacob and Monod 1961) et la découverte du code génétique, on sait que la succession des nucléotides n'est pas due au hasard mais porteuse d'une information qui se manifeste sous la forme d'un caractère (Weatherall 2001) ou d'une fonction. Or, pour Longo et coll., les phénomènes biologiques sont des phénomènes stochastiques et pour cela, ils ne peuvent être assimilés à de l'information (je reviendrai sur cette problématiques dans le second chapitre). Plus récemment, une nouvelle définition du concept de l'information a été proposée par Floridi (Floridi 2005) basée sur les données. On peut la résumer de la façon suivante :

$$\boldsymbol{Information = données + sens}$$ (Leleu-Merviel and Useille 2008).

Le sens, ici, a été défini par Rastier comme un « *phénomène contextuel* ». Cette définition a été adoptée par une large communauté scientifique (Leleu-Merviel and Useille 2008) car elle

intègre la dimension sémantique. Notons ici l'importance du contexte dans le processus informationnel.

2. Histoire de l'information

En 1943, une équipe de l'Institut Rockefeller (Avery et al. 1944) confirme que l'ADN est le support de l'information génétique. Cette expérience a été confirmée par les travaux d'une équipe du Carnegie Institution of Washington et qui a démontré le rôle de l'ADN dans l'hérédité (Hershey and Chase 1952). Ces expériences affirment que le phénomène de transmission de l'information est un phénomène qui date depuis l'apparition de la vie. L'ADN constitue la première information structurée et transmise à travers des phénomènes tels que la division cellulaire (Skarstad et al. 1983), la reproduction (Camefort et al. 2016) et la conjugaison (Holmes and Jobling 1996). Il s'agit ici d'une transmission de caractères et de modes de transmission innés. Grâce à cette information, on peut, par exemple, maintenant retracer une partie de l'histoire du peuplement humain (Cavalli-Sforza and Feldman 2003, Cherni et al. 2009, 2016, Hajjej et al. 2016). Avec l'évolution et la complexification de la vie, l'information et ces modes de transmission ont évolué. Dans le règne végétal, des mécanismes de transmission de signaux ont été découverts récemment qui peuvent transmettre un signal entre des feuilles distantes (Hedrich et al. 2016). Chez les insectes également, un mode de transmission de l'information via des médiateurs chimiques a été observé notamment chez la fourmi (Stroeymeyt et al. 2014). Chez les animaux vertébrés, l'évolution du mode de vie « la notion de vie en communauté » et « le déplacement » ont contribué à l'apparition de nouveau mode de transmission de l'information. Le système d'information est basé sur le son et l'émission de cris (Charles F. 1960, Falk 2004). Plus récemment, avec l'évolution des espèces et l'apparition des hominidés il y a plus de 2 millions d'années (Reed et al. 2004), une nouvelle notion va apparaître : la notion des connaissances et du savoir. Cette information, un peu plus complexe, a induit des formes de transmission plus évoluée. L'*homo sapiens*, l'homme moderne aurait inventé le langage (Perreault and Mathew 2012) il y a plus de 150 000 ans, instaurant ainsi une nouvelle forme de transmission de l'information. Le langage constituerait ainsi le premier formalisme pour la transmission du savoir. La transmission verbale est alors le premier système d'information inventé par l'homme pour pérenniser les connaissances acquises. Ceci représente un des premiers signes d'une intelligence supérieure. Cependant, le langage n'a pas été le seul formalisme inventé par l'homme pour la transmission d'information. La deuxième révolution a été l'invention de l'écriture. En effet, l'homme dans son évolution va chercher à matérialiser l'information acquise. Dans ce sens, l'art rupestre va être l'ancêtre de l'écriture (Masson 2006).

En effet, l'invention de l'écriture est un pas important dans la transmission de l'information. Le premier langage écrit a été daté chez les sumériens, il y a plus de 5 000 années (Gilmont 2004) sous forme de tablette en argile peu minable (voir annéxe). L'homme va adapter sa technologie au besoin de transmission de l'information en inventant le papyrus en Egypte au milieu du troisième millénaire av. J-C, puis le parchemin deux siècles av. J.-C. (Green 1993), et enfin le papier en Chine au premier siècle av. J.-C. (Needham and Tsuen-Hsuin 1985) qui va favoriser une nouvelle manière pour la transmission du savoir en produisant le livre (Figure 2). Le livre va complétement changer le paradigme de transmission et de stockage de l'information et alors permettre de consigner et de diffuser l'information de façon beaucoup plus simple et une nouvelle façon de centralisation de l'information va apparaître : les bibliothèques (la bibliothèque fût d'Alexandrie construite trois siècles av. J.-C.) (Collins 2000).

Figure 2. Le Nag Hammadi codex. **Il est connu comme étant le plus ancien livre relié au monde. Son âge est estimé à plus de 1 600 ans. Museum Copte du Caire en Egypte.**

L'analyse de l'information par des chercheurs a été initiée à l'époque de l'empire Abbaside (750-1258 apr. J.-C.) (Clark 1901). Les travaux de l'époque pour le stockage, la conservation, l'indexation, la traduction et l'analyse de l'information présentée dans les différents supports (parchemin, papiers et livres) constitueraient les premières prémisses des sciences libraires et de la bibliothéconomie, un des fondements des sciences de l'information. Si les avancées scientifiques ont contribué à l'adaptation des technologies de transmission et de conservation

de l'information, la nécessité de les transmettre a contribué à la création de nouveaux supports. En 1665, la Royal Society a édité le premier journal scientifique connu « *Philosophical Transactions* »6 (Figure 3). Aujourd'hui, on compte plus de un milliard de références scientifiques rien que dans la base de données « *Web Of Science* »[7].

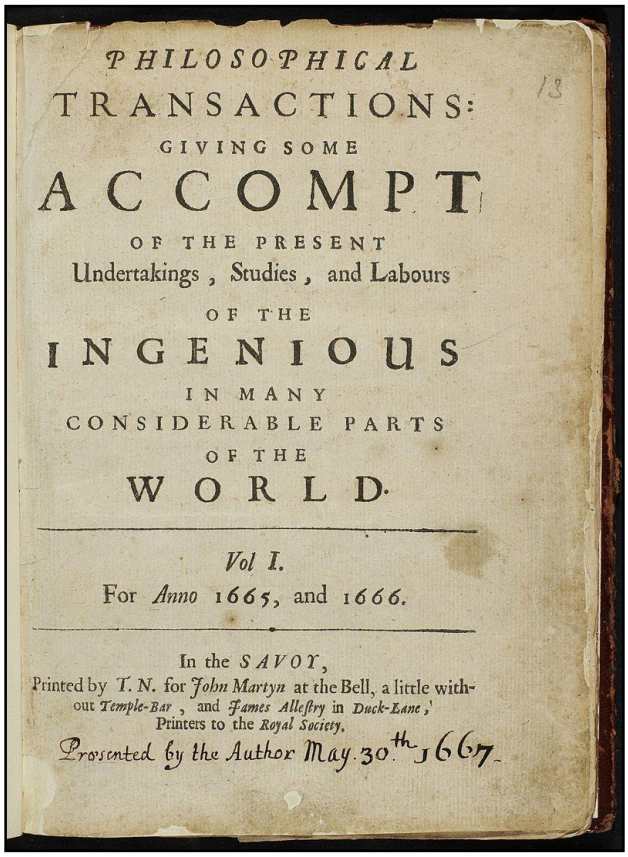

PHILOSOPHICAL
TRANSACTIONS:
GIVING SOME
ACCOMPT
OF THE PRESENT
Undertakings, Studies, and Labours
OF THE
INGENIOUS
IN MANY
CONSIDERABLE PARTS
OF THE
WORLD.
Vol I.
For Anno 1665, and 1666.
In the SAVOY,
Printed by T. N. for John Martyn at the Bell, a little without Temple-Bar, and James Allestry in Duck-Lane,
Printers to the Royal Society.
Prosented by the Author May. 30th 1667

Figure 3. La première page du premier volume de la revue Philosophical Transactions. Source : http://rstl.royalsocietypublishing.org/

L'invention de l'imprimerie par Johannes Gutenberg dans le 15ème Siècle (Zeigler 1997) et la révolution industrielle a modifié aussi les pratiques d'édition avec l'introduction de la mécanisation et la possibilité d'imprimer en masse. Cette augmentation de la production documentaire et la facilitation de la diffusion ont induit l'émergence d'une nouvelle discipline pour sa gestion. C'est ainsi que les sciences documentaires ont vu le jour au début du 20ème siècle (Otlet 1934). Cette science avait pour but l'étude des méthodes d'enregistrement et de récupération de l'information. Une autre révolution va influencer l'émergence des sciences de l'information. En 1854, George Boole présente les bases des mathématiques booléennes qui

[6] http://rstl.royalsocietypublishing.org/

[7] http://ipscience.thomsonreuters.com/product/web-of-science/

seront très utilisées par la suite pour la récupération des données (Boole 1854, Smith 1993). L'invention des première machines à calculer et des ordinateurs suite aux travaux Herman Hollerith (fondateur d'IBM) (Scientific American 1890) et après la publication des travaux de Turing sur la calculabilité (Turing 1936) et ceux de Von Neumann sur l'architecture (Von Neumann and Godfrey 1945) vont conduire à la mise au point du premier ordinateur électronique programmable pour résoudre des problèmes calculatoires (Goldstine and Goldstine 1946) à l'université de Pennsylvanie. L'introduction des transistors et surtout l'invention des microprocesseurs (Aspray 1997) vont contribuer à la démocratisation de l'informatique en réduisant la taille des ordinateurs et ainsi permettre le développement de nouveaux moyens de stockages tels que le disque dur, la disquette et le disque compact. Le développement des langages de programmation (Kowalski 1974), des bases de données relationnelles (Codd 1970) et des systèmes de gestion de base de données (Schek and Pistor 1982) vont complétement changer les pratiques de l'information. L'information commence petit à petit à se dématérialiser ; les bases de données s'avèrent être un outil puissant de stockage peu onéreux en termes de coût, d'espace et de main d'œuvre. Ceci va pousser les spécialistes à repenser leurs approches de l'information pour profiter de ce fort potentiel technologique. Durant les années 1950, la science de l'information proprement dite est née[8]. C'est Borko qui a posé la première définition de cette science (pragraphe 2 de la définition est donnée dans le cette section).

En 1969, ARPAnet, un réseau reliant à l'époque des ordinateurs de quatre centres universitaire (Université de Californie Santa Barbara, Université de Californie Los Angeles, Université de Stanford et l'Utah), est inventé (Glowniak 1998). Ce réseau a été adapté au début aux spécialistes et était soutenu par le ministère de la défense américaine[9]. En 1980, ARPAnet est divisé en deux réseaux dont le NFSnet dédié aux utilisations civiles (Glowniak 1998). En 1983, le TCP/IP (Cerf and Khan 1974), un ensemble de protocoles pour l'échange de données en réseau, a été adopté puis l'introduction, dans les années 1990, des services « user-friendly » surtout le « *world wide web* » va conduire à la naissance de l'Internet moderne (Berners-Lee, et al. 1992). Cette révolution numérique va s'accentuer avec l'apparition et la démocratisation des Smartphones surtout après le lancement de l'iPhone® en 2007[10]. En 2012, on estimait la

[8] https://www.asist.org/about/history/

[9] http://www.darpa.mil/about-us/timeline/modern-internet

[10] http://www.apple.com/pr/library/2007/06/28iPhone-Premieres-This-Friday-Night-at-Apple-Retail-Stores.html

part des smartphones à 20% du nombre total des nouvelles acquisitions de téléphones, estimé à 4,4 milliards d'unités (Teacher et al. 2013). Cette évolution technologique va accroître l'offre documentaire et les capacités de stockage (Hilbert and López 2011). La Figure 4 proposée par Gillings, Hilbet et Kemp montre l'évolution de l'offre informationnelle au fil du temps. Cette évolution va nous amener à repenser notre approche de l'information.

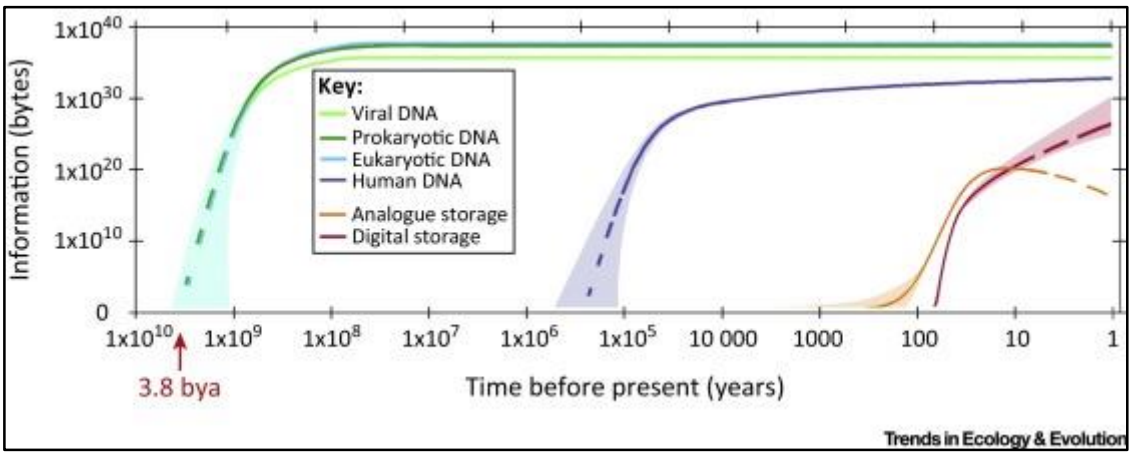


Figure 4. Illustration schématique de la quantité croissante d'informations dans la biosphère au fil du temps. D'après (Gillings, Hilbert, et Kemp 2016), source : http://www.sciencedirect.com/science/article/pii/S0169534715003249.

C'est ainsi que des concepts tels que *big data* et *cloud computing* sont apparus. Le terme *big data* va commencer à émerger vers le début années 2000 suite à l'explosion de l'offre Internet. Cette offre de services, d'outils et d'applications a facilité la production de nouvelles données et la digitalisation de données anciennes en masse (Google books [11] en est un exemple), et la diffusion des données (De Mauro et al. 2016). C'est dans ce contexte que le *big data* est né (Hilbert and López 2011, Boyd and Crawford 2012, De Mauro et al. 2016). Contrairement à ce que laisse penser le terme, on ne pourrait pas le réduire à une simple description des données massives. En effet, Boyd et Crawford (Boyd and Crawford 2012) définissent le *big data* comme un phénomène culturel, technologique et savant qui repose sur l'interaction de :

- La technologie : maximisation de la puissance de calcul et de la précision des algorithmes pour rassembler, relier, analyser et comparer de grande masse de données,
- L'analyse : plonger dans de grandes masses de données, afin d'identifier des tendances pour de nouvelles hypothèses ou des affirmations,

[11] https://books.google.com/

- La mythologie : la croyance que les données massives seraient plus intelligentes. Elles peuvent de ce point de vue générer des connaissances nouvelles plus précises et plus objectives.

Le *cloud computing* (informatique en nuage) est la nouvelle génération des centres de données avec des nœuds « virtuels » à travers des technologies d'hyperviseurs[12], comme les machines virtuelles, personnalisables et flexibles selon la demande et accessibles à travers des technologies de services web (Buyya et al. 2009). L'intérêt de cette technologie est la très grande capacité de stockage et de calcul qu'elle peut offrir mais aussi la possibilité d'adapter l'offre à la demande contrairement aux ordinateurs classiques.

J'ai exposé ici l'évolution de l'information du stade molécule avec l'apparition de la vie, jusqu'au stade byte avec la révolution numérique actuelle. Au cours de cette évolution, les moyens de transmission de l'information ont évolué. Au début, il y a eu une évolution physiologique où le mécanisme de transmission était basée sur la reproduction, jusqu'à l'apparition du langage et l'adaptation du système nerveux et du système vocal chez l'*homo sapiens* (Pisanski et al. 2016). Puis, l'homme a adapté sa technologie pour la transmission et la conservation de l'information depuis l'écriture, jusqu'aux nouvelles technologies de communication et d'information. La science de l'information est une science récente apparue au 19ème (Rayward 1997) : ses fondements ont été définis par (Borko 1968) vers la fin des années 1960.

3. Définition des sciences de l'information

Plusieurs définitions ont été proposées pour les sciences de l'information. Cette multiplication des définitions peut être expliquée notamment par l'évolution technologique et surtout dans le domaine de l'informatique, une discipline très reliée aux sciences de l'information.

La première définition a été proposée par Harold Borko, théoricien des sciences de l'information, lors d'une conférence au *Georgia Institute of Technnology* en 1961 puis reprise en 1968 comme suit :

« C'est une discipline qui étudie :

- *les propriétés et le comportement de l'information,*
- *les forces qui régissent le flux d'information*

[12] En informatique, un hyperviseur est une plate-forme de virtualisation qui permet à plusieurs systèmes d'exploitation de travailler sur une même machine physique en même temps.

- *les moyens de traitement de l'information pour une accessibilité et une facilité d'utilisation optimales.*

Elle se préoccupe de l'ensemble des connaissances relatives à la source, la collecte, l'organisation, le stockage, la récupération, l'interprétation, la transmission et l'utilisation de l'information » (Borko 1968).

Dans les années 1980, cette définition a évolué Martha Williams propose cette définition dans le bulletin de la société américaine des sciences de l'information :

> *« La science de l'information réunit et utilise les théories, principes, techniques et technologies d'une variété de disciplines pour résoudre les problèmes liés à l'information. Dans cet amalgame de disciplines liées aux sciences de l'information on peut citer les sciences cognitives, la psychologie, les mathématiques, la logique, la théorie de l'information, l'électronique, les communications, la linguistique, l'économie, les sciences de la classification, la science des systèmes, la bibliothéconomie et la science de la gestion. Toutes ces disciplines sont amenées à résoudre les problèmes liés à l'information : sa génération, son organisation, sa représentation, son traitement, sa distribution, sa communication et son utilisation »* (Williams 1988).

Dans les années 2000, une nouvelle définition a émergé, proposée par Tefko Saracevic dans l'encyclopédie des sciences libraires et d'informations. Elle définit la science de l'information comme :

> « *L'ensemble des pratiques et des sciences qui traitent de la collecte, du stockage, de la récupération et de l'utilisation efficaces de l'information. Elle s'intéresse aux informations et aux connaissances consignables, ainsi qu'aux technologies et aux services connexes qui facilitent leur gestion et leur utilisation. Plus spécifiquement, la science de l'information est un domaine de pratique professionnelle et d'enquête scientifique portant sur la communication efficace des objets d'information et des informations, en particulier les enregistrements de connaissances, chez les humains dans le contexte social, organisationnel et individuel. Le domaine de la science de l'information est la transmission de l'univers des connaissances sous forme enregistrée, centrée sur la manipulation (représentation, organisation et recherche) de l'information plutôt que sur la connaissance de l'information*» (Bates and Maack 2009).

Ces définitions ont été adoptées par la société américaine des sciences et technologies (ASIS&T) de l'information. Néanmoins, il ne faut pas confondre science de l'information et théorie de l'information. En effet, la théorie de l'information est un concept mathématique qui permet de quantifier l'information et sa transmission. Il ne faut pas confondre aussi sciences de l'information et sciences bibliothécaires. Si la bibliothéconomie est l'ancêtre des sciences de l'information, le contexte historique n'est plus le même. En effet, la bibliothéconomie s'intéresse à l'organisation du savoir et à la gestion des documents (Constantin and Richter

2006). L'émergence de l'informatique et des outils numériques surtout la démocratisation de l'Internet et la dématérialisation de l'information a bouleversé les pratiques et les méthodes de management de l'information et a favorisé l'émergence des sciences de l'information.

4. Paradigme des sciences de l'information

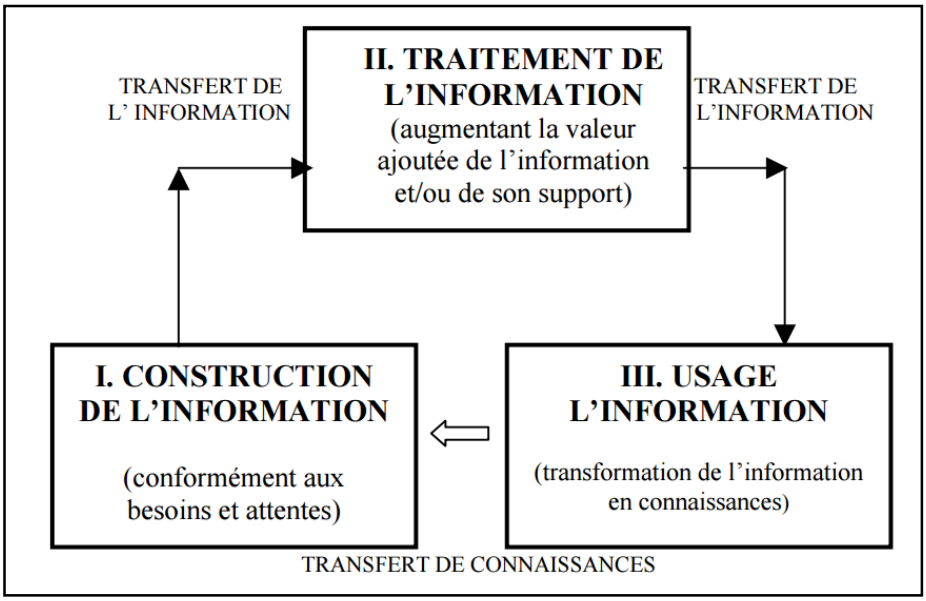


Figure 5. Usage de l'information. **Modèle à trois processus proposé par (Dragulanescu 2003).**

Selon le modèle classique proposé par Dragulanescu (Dragulanescu 2003), les sciences de l'information sont fondées sur trois processus fondamentaux. Ces processus se succèdent d'une manière cyclique (Figure 5) :

I. Construction de l'information : générer les informations à partir d'événements ou de savoirs connus et adaptation du support de présentation,

II. Traitement de l'information : via des méthodologies d'organisation, de représentation et de stockage pour augmenter la valeur ajoutée de l'information traitée et un confort d'utilisation pour l'usager,

III. Usage de l'information : diffusion et exploitation de l'information pour la transformer en connaissance.

De ce paradigme, on peut conclure une nouvelle définition de la science de l'information : il s'agit de l'étude des techniques par lesquelles on stocke, on présente et on explore l'information. Cette notion d'exploration des données (« de l'information ancienne ») pour la production de nouvelles connaissances fait appel à des techniques informatiques et statistiques

qui seraient -selon plusieurs auteurs- du domaine des sciences des données (« *data science* ») (Provost and Fawcett 2013, Cao 2016). En effet à l'ère du « Big Data » les deux disciplines sont devenues indissociables. On assiste aujourd'hui, vue l'évolution des pratiques et des technologies à une explosion des données (Partyko 2009). Ces données sont rendues de plus en plus disponibles. Par exemple l'émergence du *cloud computing* et de l'utilisation des réseaux sociaux ont contribué à l'apparition de nouvelles pratiques de commercialisation telle que le « Web Marketing » et le « Marketing Ciblé ». Ceci passe par un processus d'analyse des données collectées relatives aux habitudes du consommateur sur les web « la listes des sites visités », « l'analyse des commentaires et des messages postés sur les réseaux sociaux », les listes d'achat. Ces données constituent une mine d'or pour les entreprises qui cherchent de plus en plus à produire des offres personnalisées pour les consommateurs (Bodier et Guerout 2017). Le principe de la démarche est de partir des données collectées, non structurées, pour produire de l'information structurée qui sera par la suite analysée par des approches de fouilles des données pour produire une connaissance sur le consommateur (Partyko 2009). Ces approches qui ont soulevé quelques questionnements d'un point de vue éthique deviennent de plus en plus utilisées en économie moderne, en témoigne le développement exponentiel des entreprises détentrices de ces données (Rameaux 2017 ; Nikos 2016 ; Smyrnaios et Franck 2009) communément nommées les GAFAM[13]. Ces approches pour la production de nouvelles connaissances à partir de données non structurées vont être aussi utilisées pour des fins scientifiques. En effet, comme on va le voir plus tard en biologie, domaine qui a vu un développement considérable de données expérimentales, notamment suite à la mise en place des technologies à haut débit. Cette augmentation exponentielle de la masse des données générées s'est accompagnée de l'émergence de la bioinformatique comme discipline qui s'intéresse au traitement et à l'analyse des données biologiques. Pour résumer nous pouvons décrire notre ère comme l'ère des données massives. Ces données sont générées de manière non structurée d'où l'intérêt de les structurer et de les analyser pour produire de la connaissance et stocker les nouvelles informations ainsi générées. Pour cette raison nous proposons un nouveau schéma inspiré du paradigme de Dragulanescu (Dragulanescu 2003) dans la Figure 6, où nous

[13] GAFAM, est un nom générique qui a été données aux géants américains du numérique, c'est l'abréviation des 5 entreprises américaines Google (Moteur de recherche, solution de courrier électronique, et cloud computing), Amazon (Commerce en ligne et cloud computing), Facebook (Réseaux social le plus utilisé), Apple (constructeur d'ordinateurs, de téléphonnes mobiles et tablettes) et Microsoft (Logiciels informatiques, courrier électronique, constructeurs de téléphones et de d'ordinateurs, cloud computing).

décomposons l'étape de structuration des données en deux parties aynats des objectifs et des méthodes de mise en oeuvre différents.

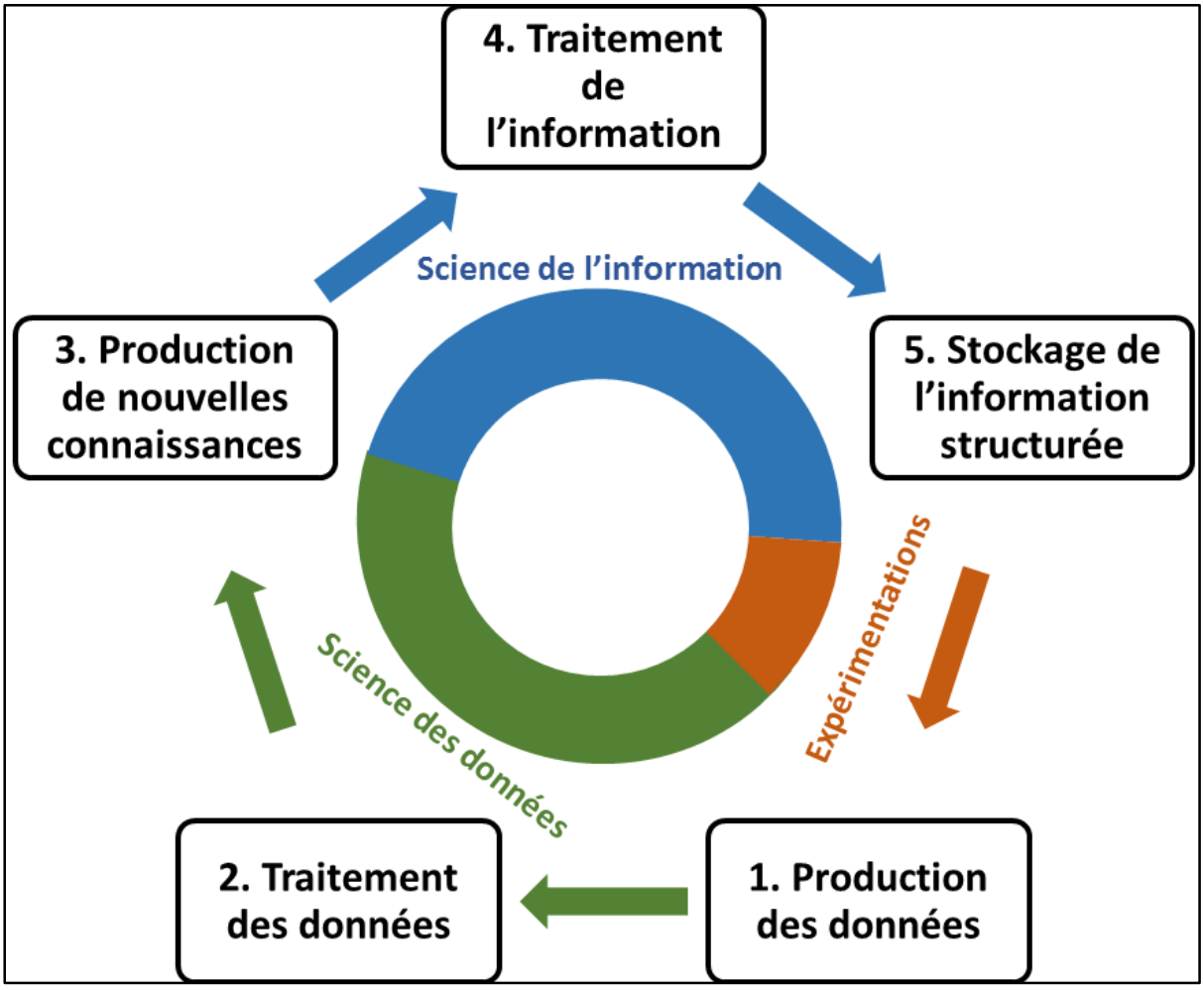


Figure 6. Nouveau paradigme de l'information. Ici on passe de la donnée non structurée générée par l'expérimentation (1) en utilisant des approches algorithmiques des sciences des données on y décèle de la connaissance (2-3). Cette nouvelle connaissances sera structurée contextualisée pour avoir un sens et devenir une information(4). Source : Selon Dragulanescu (Dragulanescu 2003) avec modifications.

5. Méthodologies, approches et technologies des sciences de l'information

Dans cette partie j'exposerais des approches informatiques utilisée en science de l'information pour le traitement, la structuration et le stockage de l'information pour la rendre disponible.

5.1. Les systèmes d'information

Un système d'information est un système qui permet de collecter, stocker, traiter et distribuer l'information dans un environnement donné (De Courcy 1992) informatiquement. C'est un système socio-technologique dont la composante sociale comprend la structure

organisationnelle en plus des différentes entités qui composent le système, et la composante technologique comprend la partie hard et software pour le développement du système. Par exemple en informatique les bases de données, et les portails Internet sont des systèmes d'informations très utilisés. En management, les progiciels de gestion intégrés sont des systèmes d'information qui permettent de centraliser les données pour la gestion des ressources humaines, financières et matérielles pour améliorer la gestion. Les systèmes d'information géographiques (SIG) par exemple ont des applications qui vont de la santé (Molla et al. 2017) à l'économie et l'environnement (Serbu et al. 2016). Ce type particulier de système d'information a été conçu pour recueillir, stocker, gérer et présenter tous les informations de type spatiales et géographiques. Dans le domaine de la santé une revue publiée en 2014 sur 51 51 SIG montre que les axes principaux d'application dans le domaine de la santé peuvent être résumés de la façon suivante : la surveillance des maladies, le soutien de aux systèmes de santé, la promotion de la santé et la prévention des maladies, et la communication entre les prestataires de soins de santé (Nhavoto et Grönlund 2014).

Ces systèmes d'informations utilisent le plus souvent plusieurs sources d'informations d'où le besoin de standardisation et d'homogénéisation de l'information, on va présenter dans ce qui suit une aperçu sur les efforts qui ont été entrepris pour atteindre ce but.

5.2. Ontologie

La première approche de standardisation que j'expose ici est les ontologies.

5.2.1. Définition d'une ontologie

> « *Ce terme trouve sa racine dans la philosophie classique. Il a été utilisé au début par le philosophe Aristote dans sa « métaphysique »*[14]*. C'est une branche de la philosophie qui traite de la nature et la structure de la réalité* » (Guarino et al. 2009).

À l'instar des sciences expérimentales, qui visent à découvrir et à modéliser la réalité sous une certaine perspective, l'ontologie se concentre sur la nature et la structure des choses, indépendamment de toute autre considération et même indépendamment de leur existence réelle (Guarino et al. 2009). Le terme a été depuis repris par les informaticiens et les spécialistes de l'information. En effet, le concept a été introduit pour résoudre des problèmes liés à l'extraction des informations, sa représentation et l'organisation des systèmes de connaissances (Xuning et al. 2012, Ivanović and Budimac 2014). Il n'y a pas une définition unanime des informaticiens

[14] Dans ses traités de la logique connue sous le nom d'organon, Aristote traite de la nature du monde. La métaphysique était le sujet prédominant dont l'Ontologie. Plusieurs philosophes ont confondu alors métaphysique et ontologie.

pour ce concept. Par contre beaucoup d'auteurs (Smith 2003, Guarino et al. 2009, Xuning et al. 2012, Ivanović and Budimac 2014) citent la définition suivante donnée par Gruber (Gruber 1993) « une ontologie est une spécification explicite d'une conceptualisation ». En 1997, Borst (Borst 1997) propose la définition suivante : « une ontologie est une spécification formelle d'une conceptualisation partagée ». Les deux définitions utilisent le terme « conceptualisation ». Nilsson (Nilsson 1991) va définir la conceptualisation comme un ensemble de concepts (entités) reliés entre eux. Plus tard, Gruber (Gruber 1995) va rajouter la notion d'abstraction (Guarino et al. 2009). Par conséquent, selon la première définition, une ontologie est une modélisation conceptuelle d'un domaine de connaissance. Borst (Borst 1997) inclut la notion de formalisation donc de standardisation et de partage. Pour conclure, une ontologie est un vocabulaire standard pour modéliser un domaine de connaissances donné. Ce vocabulaire, par définition, doit être partagé par les experts du domaine. Sur un plan fonctionnel, une ontologie est organisée de manière hiérarchique, du terme le plus général (racine) vers les termes les plus spécifiques, sous la forme d'un graphe acyclique orienté (Dutkowski et al. 2014). C'est sur cette base que les ontologies biologiques ont été développées (Ashburner et al. 2000, Bard et al. 2005, Mungall et al. 2012, Kibbe et al. 2014, Groza et al. 2015).

Dans la Figure 7 est présenté un fragment de la *Cell Ontology*, une ontologie des populations cellulaires. Cette ontologie a pour but de répertorier toutes les populations et sous populations cellulaires (des procaryotes aux mammifères)[15], de décrire les processus biologiques qui y sont associés ansi que certaines localisations anatomiques. La première étape consite à developper une nomemclature standardisée pour la description des « types cellulaires » ansi que leurs définitions, les synonymes utilisés dans littérature et les propriétés. Ainsi on parlera de concepts cellulaires. Certains types cellulaires partagent des fonctions biologiques, agissent dans les mêmes voies biologiques. Ils peuvent aussi partager certains récepteurs membranaires ou mêmes des caractéristiques transcriptomiques. Ils peuvent aussi appartenir aux mêmes systèmes biologiques ou aux mêmes tissus. Des types cellulaires peuvent aussi provenir de l'évolution d'autres types par différenciation. Ces caractéristiques ont été utilisées pour regrouper ces types cellulaires de manière hiérarchique, on parle alors de notion de populations et de sous populations cellulaires. Pour les localisations anatomiques et pour les processus biologiques qui impliquent certaines populations, *Cell Ontology* a été enrichie à partir du vocabulaire d'autres

[15] http://obofoundry.org/ontology/cl.html

ontologies telles que *UBERON ontology* (pour les localisations anatomiques) et *Gene Ontology* (pour les processus biologiques) (Diehl et al. 2016).

5.1.1. Les ontologies, pourquoi ?

Le principal objectif derrière le développement des ontologies est de pouvoir construire des systèmes informatiques qui peuvent échanger entre eux et échanger des informations avec les humains assez facilement (Ivanović and Budimac 2014). Le développement de tels systèmes implique l'utilisation d'informations de sources hétérogènes et donc nécessite l'utilisation d'une terminologie standard. Plusieurs intervenants peuvent être impliqués pour le développement et l'utilisation de ces systèmes ; cette terminologie doit donc être partagée. Ce système intelligent doit aussi, avoir une capacité de « raisonnement » et la capacité de lier les sources d'information. Les connaissances utilisées et générées par ce type de système doivent être réutilisables et interopérables. Les ontologies par leurs structures en graphe et la nature du vocabulaire utilisé sont développées en ce sens (Chandrasekaran et al. 1999).

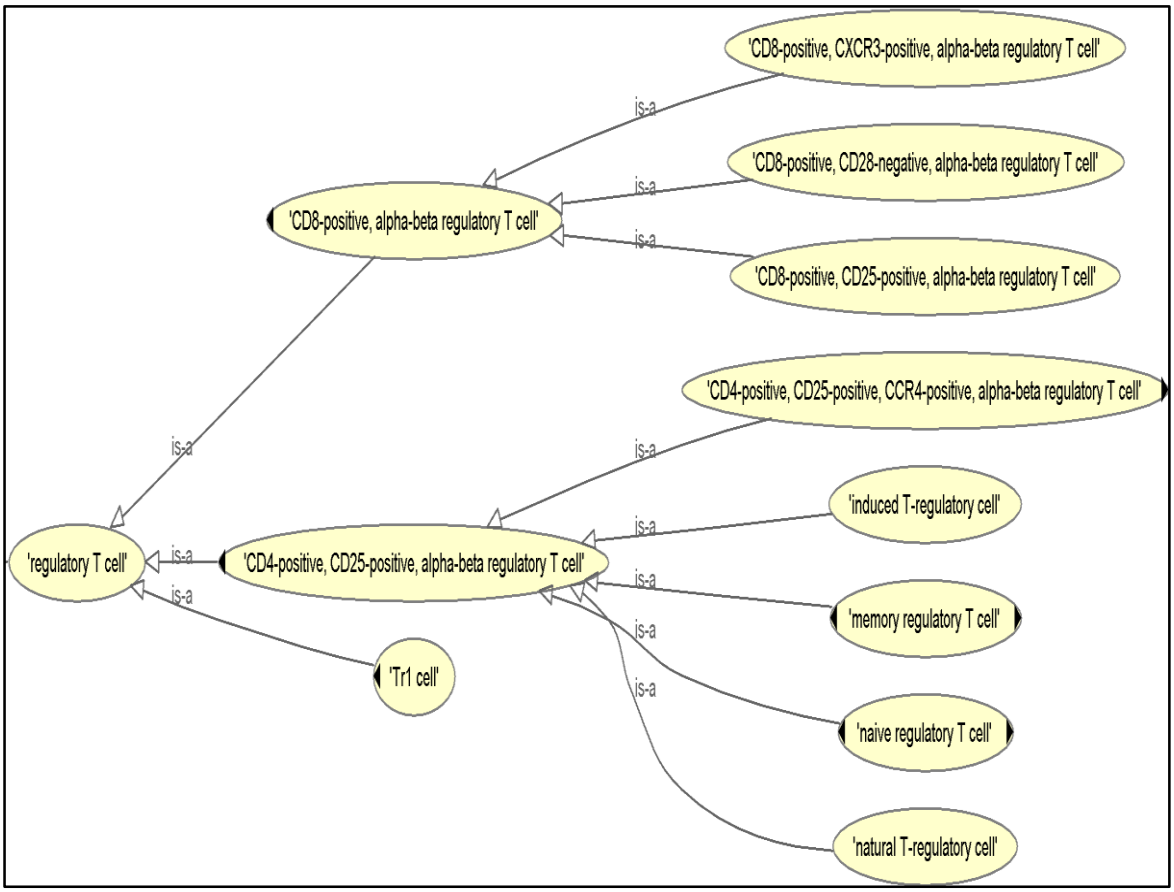


Figure 7. Représentation d'un fragment de l'ontologie Cell Ontology. Ce fragment a été généré par l'éditeur d'ontologie *Protégé*. Ici sont représentés les concepts des sous populations des cellules T régulatrices (« *Regulatory T Cell* » qui se trouve en haut, une population qui subit des altérations lors du vieillissement du système immunitaire et des maladies auto-immunes). Les liens de parentés sont les fléches resprésentées par la relation « is a » qui partent des termes fils (sous population) vers les termes parents.

5.2. Le Web Sémantique

Le web sémantique est un concept qui a été présenté par Berners-Lee et coll. (Berners-Lee et al. 2001), le fondateur du web classique. Le but étant de rendre « le web plus intelligent ». Le web classique a été conçu pour être lu par des humains et pour leur permettre de naviguer de page en page via les liens hypertexte (Berners-Lee, et al. 1992). Alors que la mission du web sémantique est de créer une toile de données « intelligentes » qui peuvent être traitées par des machines via des technologies d'intelligence artificielle pour améliorer l'accès à l'information (Berners-Lee et al. 2001, Shadbolt et al. 2006). Le langage HTML classique permet une description informatique classique du contenu d'une page Web et des liens. Avec le web sémantique, il s'agit de rajouter des annotations sémantiques au contenu, ce qui va rajouter la notion du sens à la page Web et donc on va décrire non seulement le contenu mais aussi la structure des connaissances que la page contient. Ainsi une machine peut :

- Utiliser des procédés tels que les raisonneurs (Sirin et al. 2007) pour traiter de la connaissance elle-même au lieu du texte.
- Ceci peut améliorer la qualité de la recherche et faciliter la collecte automatisée d'informations pour obtenir des résultats plus significatifs

Se pose alors un problème d'interopérabilité[16] des données disponibles. En effet, c'est la contrainte principale pour le développement de ces outils d'intelligence artificielle. L'abondance des données, de leurs sources et des domaines sur Internet induit une grande variété des terminologies présentes dans les bases de données (Berners-Lee et al. 2001, Shadbolt et al. 2006, Ivanović and Budimac 2014) et pose un problème pour la communication entre systèmes. Une partie de la réponse a été apportée par le développement des ontologies qui constituent un vocabulaire standard (Berners-Lee et al. 2001) composé de concepts liés sémantiquement. Donc des langages de programmation et des formalismes spécifiques ont été développés pour permettre l'intégration des annotations sémantiques telles l'OWL (Ontology Web Language) (Bechhofer 2009) et le RDF (Resource Description Framework)[17].

[16] Capacité de matériels, de logiciels ou de protocoles différents à fonctionner ensemble et à partager des informations. Source : http://www.larousse.fr/dictionnaires/francais/interop%C3%A9rabilit%C3%A9/43787

[17] https://www.w3.org/TR/2004/REC-rdf-concepts-20040210/#section-Concepts

5.3. Les linked data (Web des données)

Linked Data consiste simplement à utiliser le Web pour créer des liens typés entre les données provenant de différentes sources. Ceux-ci peuvent être des bases de données provenant de différentes organisations dans différentes localités géographiques, ou simplement des systèmes hétérogènes provenant d'une seule organisation mais qui ne sont pas interopérables au niveau des données. Techniquement, les *Linked Data* sont des données publiées sur le Web de manière à être aussi lues par les machines ; elles sont liées à d'autres ensembles de données externes (Bizer et al. 2009).

Linked Data repose sur deux technologies les :

- URI[18] : (*Uniform Resource Identifiers*) identifiant uniforme de ressource. C'est un système plus générique que les URL mis en place par le consortium « *World Wide Web* » (W3C) pour identifier n'importe quel objet sur internet dans le monde.
- http[19] : (*HyperText Transfer Protocol*) est un protocole de communication, client-serveur développé par le W3C.

- **Projet lié (Open Linking Data)**

Le projet *Open Linking Data* est un projet qui a été lancé en 2007 (Bizer, Heath, et Berners-Lee 2009), et qui était soutenu par le groupe « *Semantic Web Education and Outreach* » (SWEO)[20] du W3C. Le but étant d'effectuer un « bootstrap » des données du web pour identifier des ensembles de données publiques, de les convertir en suivant les recommandations RDF et linked data, et de les republier sur le Web (Bizer et al. 2009). Les premiers participants au projet étaient au début des universitaires, des petites entreprises et des particuliers. Vu la nature du projet, il a connu une croissance rapide, car tout le monde peut publier des données en suivant les principes du Linked Data. La Figure 8 et la Table 1 présentent les données disponibles et les disciplines concernées.

[18] https://tools.ietf.org/html/rfc3986

[19] https://www.w3.org/Protocols/HTTP/1.1/rfc2616.pdf

[20] https://www.w3.org/blog/SWEO/

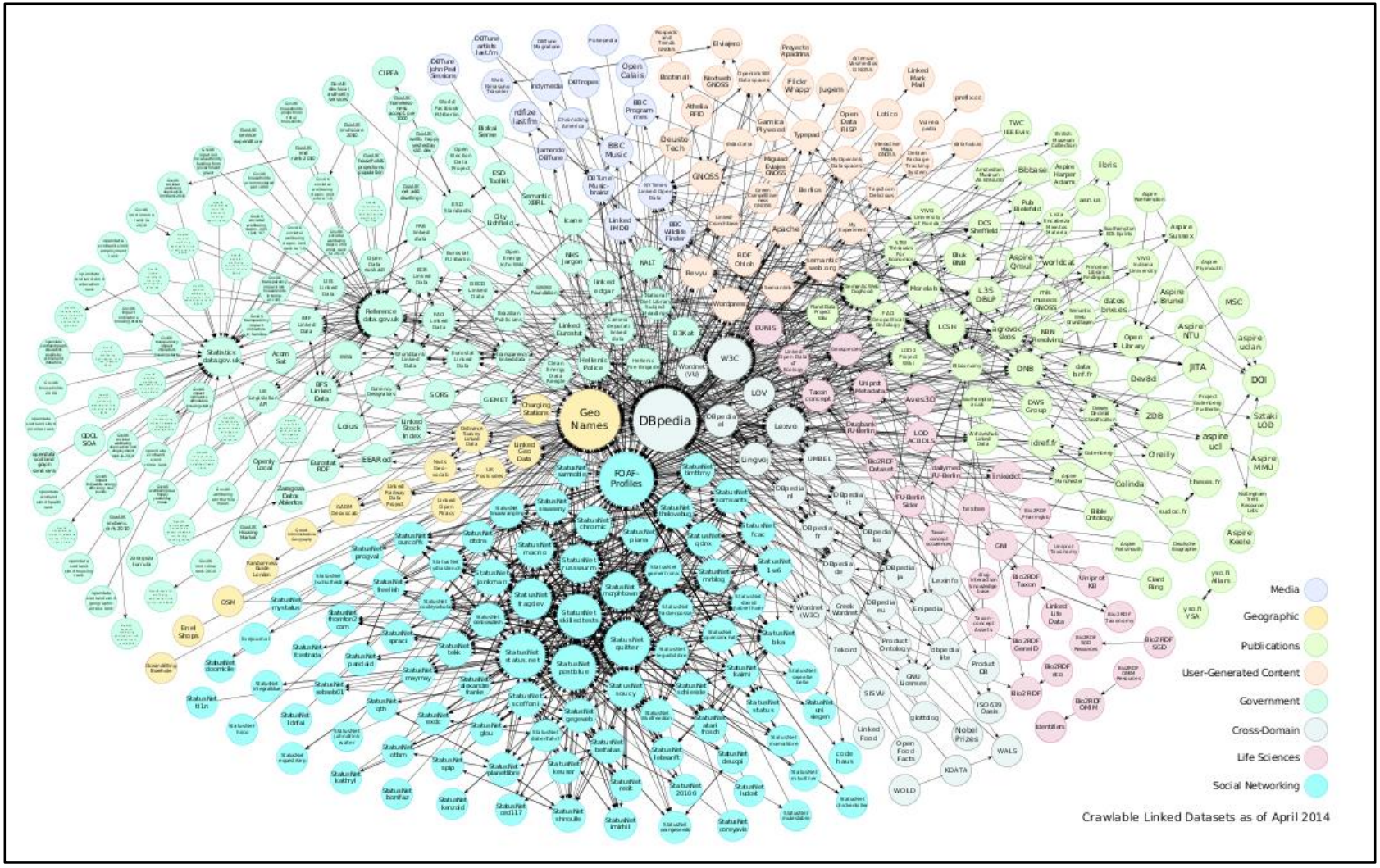


Figure 8. Diagramme en nuage donnant un aperçu sur les ensembles de données liées dans le cadre du projet Linked Open Data. Le diagramme a été produit durant le mois d'avril 2014. Les cercles en mauve sont des données liées aux sciences de la vie. Source : http://linkeddatacatalog.dws.informatik.uni-mannheim.de/state/.

Table 1. Présentation des ensembles de données disponibles dans le cadre du projet Open Linking Data par domaine. Le terme « ensemble de données » se réfère au terme anglais « *dataset* ». Source : http://linkeddatacatalog.dws.informatik.uni-mannheim.de/state/

Domaine	Nombre de données par sources	%
Gouvernement	183	18,05%
Publications	96	9,47%
Sciences de la vie	83	8,19%
Contenue généré par les utilisateurs	48	4,73%
Inter-domaines	41	4,04%
Médias	22	2,17%
Géographie	21	2,07%
Web social	520	51,28%
Total	1014	

6. Conclusion sur les sciences de l'information

On peut conclure que l'afflux massif d'informations auquel nous assistons, dû au développement scientifique et technologique, nous met face à plusieurs enjeux. Si des solutions techniques ont été développées, ou sont en cours de développement, pour le stockage et pour rendre l'information disponible, la plus grande problématique reste la standardisation du vocabulaire et des formats. Ceci est indispensable pour faciliter les échanges entre humains (vue l'émergence de la pluridisciplinarité) (Attwood et al. 2011), et entre l'homme et la machine. Un autre enjeu est posé par le développement de l'offre informationnelle : il s'agit de l'analyse des données (de l'information). Cette analyse doit passer par des processus automatiques étant donné la masse d'information. Le développement de tels systèmes pose le problème d'interopérabilité des données. Le développement d'ontologie, du web sémantique et l'émergence de projets tels que l'*Open Linking Data Project* sont des initiatives pour la standardisation. Les sciences biomédicales ont aussi été incluses dans ces projets de standardisation des terminologies (Ivanović and Budimac 2014), rendant possible le développement de systèmes d'analyse de l'information biomédicale.

B. L'après information

D'après ce qui précède, on peut résumer les objectifs des sciences de l'information, à la structuration, la description, et la représentation de l'information. L'étape suivante consiste à

analyser et/ou classifier ces données afin de produire des nouvelles connaissances, ou bien pour prévenir la survenue de nouveaux phénomènes. La discipline scientifique pour assurer cet objectif est la science des données (Cleveland 2001). Cleveland décrit la science des données comme une discipline statistique. Néanmoins, ce paradigme est très réducteur. En effet, le même auteur plaidait dans le même article pour que l'analyse des données ou la science des données s'ouvre sur d'autres disciplines. Récemment, Cao (Cao 2016) propose une définition plus large et plus réaliste des sciences des données :

> « *C'est une science pluridisciplinaire qui se base sur les statistiques, l'informatique, la communication, la gestion la sociologie pour étudier les données et leurs environnements (y compris les aspects contextuels, tels que les aspects organisationnels et sociaux) dans le but de transformer les données en idées et en décisions en suivant la méthodologie de la donnée à la connaissance (savoir) à la sagesse* ».

Cette définition bien qu'exhaustive oublie les mathématiques alors que plusieurs auteurs utilisent des approches mathématiques et de modélisation pour l'analyse des données (Agresti and Kateri 2011). Nous pouvons ainsi définir la science des données comme une science pluridisciplinaire qui s'intéresse à la quantification, la classification et la description des données dans le but de produire une nouvelle connaissance (soit des nouvelles informations soit la prédiction d'un nouveau comportement ou phénomènes). Certains auteurs confondent science des données et science de l'information dans le sens où ils présentent la prédiction et la classification comme un paradigme et une finalité des sciences de l'information (Dragulanescu 2003).

1. Fouille des données (*Data Mining*)

1.1. Définition

Selon Han et coll. (Han et al. 2012), la fouille des données est un sujet pluridisciplinaire qui pourrait être défini de différentes façons et le terme « fouille des données », lui-même est un terme réducteur. Selon les mêmes auteurs, cette discipline aurait pu s'appeler « exploration des connaissances à partir des données ». Par analogie à l'extraction de l'or de la roche, il s'agit ici d'extraire des pépites de connaissances à partir d'une grande masse de données. D'autres termes ont été utilisés pour décrire la fouille des données : l'extraction de connaissances à partir de données, l'extraction de connaissances, l'analyse des données et/ou motifs, l'archéologie des données et le dragage de données. Néanmoins, beaucoup d'auteurs la confondent souvent avec l'extraction des connaissances des bases de données (*knowledge discovery in databases-KDD*) (Fayyad and Stolorz 1997, Abdul Rahman et al. 2016). En effet le KDD a été définit depuis les

années 1980 comme étant un ensemble de pratiques qui permettent d'extraire des connaissances à partir des données stocké dans les bases de données (Fayyad, Piatetsky-Shapiro, et Smyth 1996). Pour Han et coll. (Han et al. 2012), la fouille des données est le processus de découverte de motifs (*patterns*) et de connaissances à partir d'une grande masse de données. Ce processus est une étape d'une démarche plus exhaustive qui comprend : le nettoyage, l'intégration, la sélection et la transformation des données, puis la fouille des données, suivi de l'évaluation et de la représentation des nouvelles connaissances. Cette démarche serait le KDD proprement dit et de ce fait le datamining n'en serait qu'une étape.

D'un autre point de vue, le texte constitue aussi un support important de l'information historiquement et jusqu'à maintenant. Une grande partie de l'information scientifique par exemple se trouve dans les textes : articles scientifiques, livres, rapport et compte-rendu. Les publications scientifiques ont connu une croissance exponentielle depuis l'ère de la numérisation Figure 9. Or la grande difficulté réside dans le fait que l'information contenue dans les textes est une information non structurée (contrairement aux bases de données).

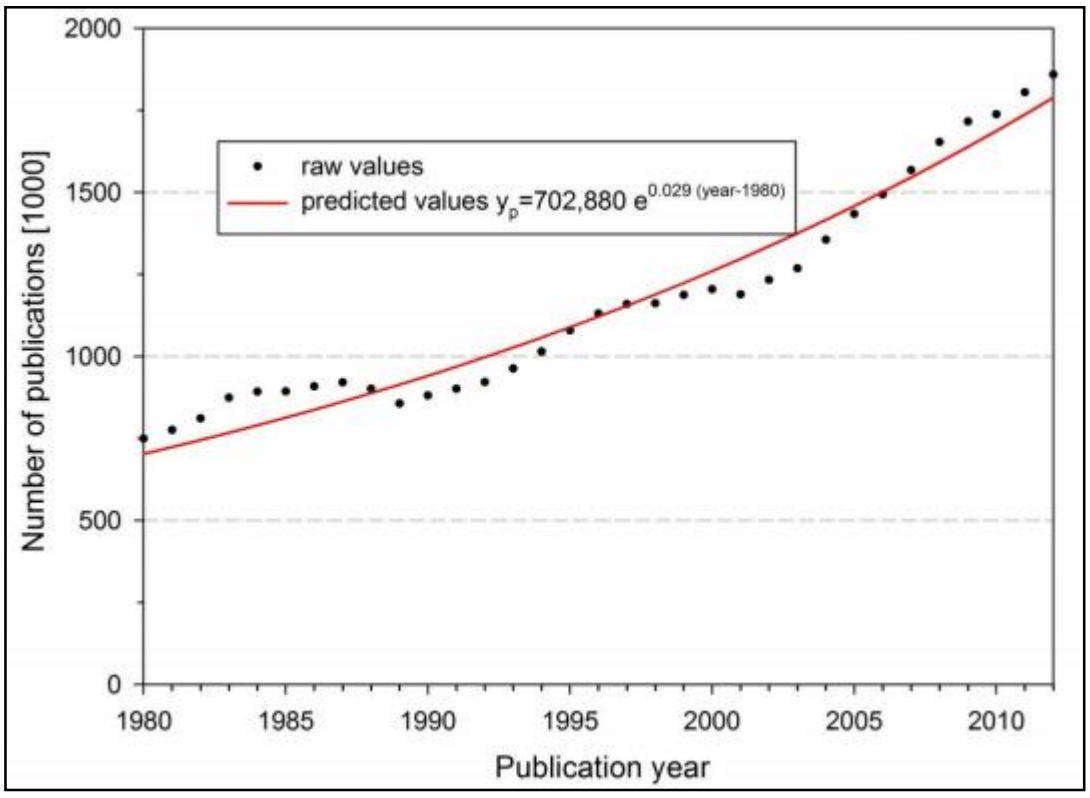


Figure 9. Croissance exponentielle de la production scientifique de 1980 à 2012. En rouge, un modèle produit par régression exponentielle ; les points noirs constituent les valeurs réelles par année. Source : (Bornmann and Mutz 2015).

La fouille des textes ou bien la fouille des textes est une branche de la fouille des données. Cette discipline utilise des approches de fouille des données combinées à des approches linguistiques (pour tenir compte de la spécificité des propriétés de cette information) pour l'extraction des connaissances à partir des textes. Selon Feldman and Sanger (Feldman and Sanger 2007), la fouille des textes est un domaine de recherche en informatique qui tente de résoudre la surcharge d'informations en combinant les techniques de fouille des données, d'apprentissage automatique, le traitement du langage naturel, la récupération de l'information et la gestion des connaissances.

1.2. Les outils de fouille des données

1.2.1. Outils statistiques

Historiquement, la collecte des données statistiques est très ancienne. En 2238 av. J.-C., l'empereur Yao recensait les produits agricoles de l'empire, les pharaons organisaient depuis le 7ème siècle av. J.-C. des recensements de la population pour la collecte des impôts (Oriol 2007). Cependant, ce n'est qu'au court du 18ème siècle que Bayes démontre qu'on peut prédire des probabilités de survenue d'un événement en se basant sur les observations d'une expérience. En d'autres termes, on peut quantifier la probabilité de survenue d'un événement a posteriori sachant la probabilité d'un événement a priori (Bayes and Price 1763). La formule de la probabilité s'écrit de la manière suivante :

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$$

$$<=> P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Avec A et B deux événements liés

P(A) et P(B) les probabilités marginales de A et B

P(A∩B) : La probabilité que les deux événement se produisent

P(A|B) : La probabilité a postériori de A sachant B

Plus tard, au début du 20ème siècle, Fisher met au point son analyse de la variance plus communément appelé l'ANOVA (Fisher 1918). Ces deux assomptions mathématiques vont être à l'origine d'une discipline statistique nouvelle, l'inférence statistique[21]. En fouille des données,

[21] L'inférence statistique est le processus de déduction et de vérification des hypothèses sur les propriétés d'une population en se basant sur les caractéristiques d'un échantillon (Larose 2005).

on va utiliser les propriétés et les méthodes d'estimation et de prédiction de cette discipline (Larose 2005). L'inférence statistique offre une panoplie de méthodes dont :

i. L'analyse par régression

Historiquement, le terme régression a été introduit par l'anthropologue britannique Galton. Il a publié en 1886 l'analyse d'une étude qu'il a menée sur une population de des cendants de personnes de grande taille (Galton 1886). Il a trouvé que la taille diminue ('régresse') de génération en génération (plaque 9 de l'article).

La régression en tant qu'analyse statistique consiste à trouver un modèle mathématique y=f(x) qui pourrait expliquer la distribution d'une variable y sachant x. Si on arrive à ajuster le modèle sur les valeurs expérimentales, on pourra par la suite l'utiliser pour la prédiction. Plusieurs modèles ont été proposés :

- La régression linaire simple :

  Le formalisme mathématique est le suivant (Weisberg 2005) :

$$y_i = \beta_0 + \beta_1 x_i$$

Les applications en science vont des courbes d'étalonnage à la classification si on utilise des données qualitatives. En macro-économie, la loi d'Okun décrit une relation linéaire (Figure 10) entre le taux de croissance d'un pays et son taux de chômage (Okun 1962). L'énoncé de la loi est le suivant :

$$TxCr = k - c * \Delta TxCh$$

Avec : TxCr, est le taux de croissance annuel du PIB

TxCh, est le taux de croissance annuel du chômage

c, est le facteur reliant les variations du chômage aux variations de la production

k, est le taux de croissance annuel moyen de la production de plein emploi

Malgré quelques critiques émises par Prachowny (Prachowny 1993) qui a réussi à démontrer que le coefficient d'Okun (c dans la formule) est variable selon le temps et selon le pays (Jim 2000). Aussi dans une étude du marché du travail égyptien, Elshamy (Elshamy 2013) démontre que le modèle ne serait valide pour les pays où le chômage est un chômage structurel. Malgré la critique sur la variabilité des paramètres, Jim (Jim 2000) insiste sur la robustesse du modèle. Plusieurs auteurs l'utilisent pour l'étude des marchés émergents tels que le Nigeria (Ola-David et al. 2016), et la Chine (Huang and Fidrmuc 2016) et les marchés traditionnels tels

que le marché européen (Dunsch 2016). Plusieurs institutions l'ont adopté pour la prédiction des taux de chômage comme la Banque Mondiale (Hanusch 2012).

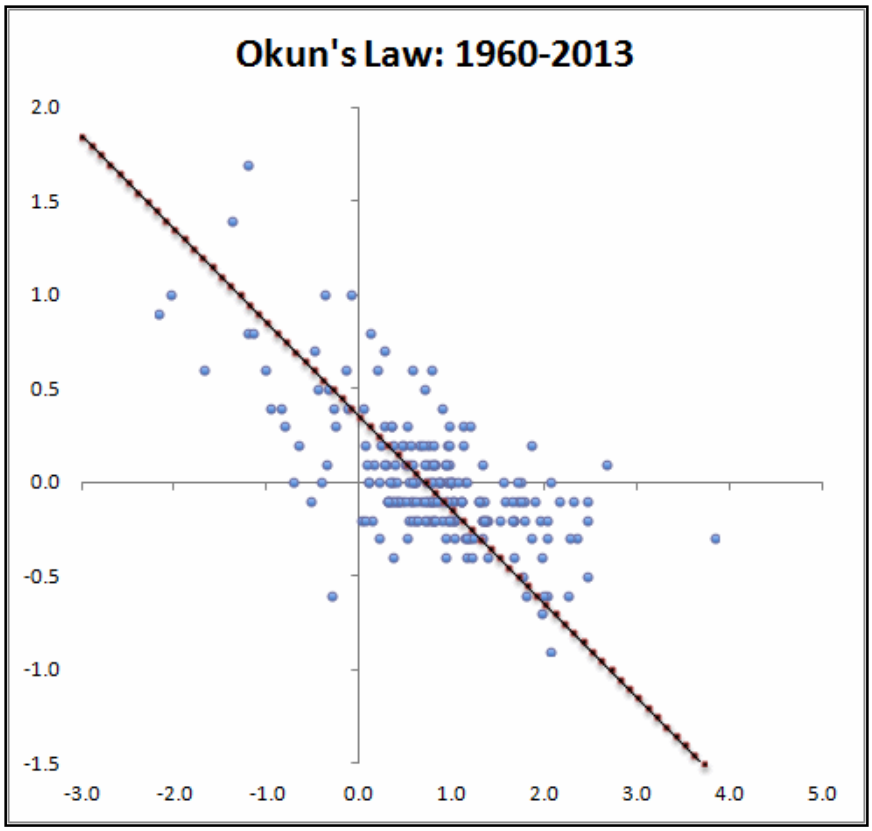


Figure 10. Illustration de la loi d'Okun pour les Etats-Unis entre 1960 et 2013. **Le taux de croissance du PIB (Horizontalement) et le taux de croissance du chômage (verticalement) sont calculée par trimestre. La relation de régression est modélisée par la courbe rouge qui permet de généraliser la relation. Source : http://www.cbsnews.com/news/explainer-okuns-law-relationship-between-gdp-and-jobs/**

- La régression linaire multiple :

La régression linéaire multiple est une généralisation de la régression linéaire simple (Eberly 2007). Le formalisme mathématique est le suivant :

$$Y_i = a_0 + a_1 X_{i1} + a_2 X_{i2} + \cdots + a_p X_{ip} + \varepsilon_i$$

Avec : Y, la variable endogène à expliquer

Xi, les variables explicatives

ε, l'erreur du modèle (elle est due au manque d'informations)

- La régression logistique :

Appelé aussi modèle LOGIT. Le formalisme mathématique est le suivant :

$$ln\frac{\pi(X)}{1-\pi(X)} = b_0 + b_1x_1 + \cdots + b_jx_j$$

Avec : π(X) traduit la probabilité de survenue de l'événement X

x_i sont les variables explicatives

Historiquement, les modèles LOGIT ont été décrits la première fois par Joseph Berkson en 1944 (Berkson 1944) pour décrire des réactions de biocatalyse. La transformation mathématique de cette relation donne :

$$\pi(X) = \frac{e^{b_0+b_1x_1+\cdots+b_jx_j}}{1+e^{b_0+b_1x_1+\cdots+b_jx_j}}$$

L'application de ces deux relations à la réaction d'hydrolyse du *sucrose* en *glucose* et *fructose* donne la relation décrite dans la Figure 11. Pour la rédaction de mon rapport je n'ai cité que les modèles de régression linéaire qui ont été utilisés dans certaine applications notement pour la modélisation des voies biologiques (voir chapitre 3), mais d'autres modèles de régressions ont été développés notamment la régression polynomiale, circulaire, ellipsoïde. Dans cette partie nous avons choisi de développer la régression comme exemple d'outil d'apprentissage statistique car il fait partie des outils les plus utilisés en biologie mais d'autre méthodes sont utilisée. Dans le paragraphe suivant nous allons citer des exemples.

Dans leur livre « Introduction to Statistical Relational Learning » Getoor et Taskar (Getoor et Taskar 2007) introduisent les modèles probabilistes graphiques. Les deux exemples les plus connus sont les réseaux Bayésiens et les chaines de Markov. Un modèle graphique probabiliste définit une famille de distributions de probabilités qui peuvent être représentées en termes de graphique. Les nœuds dans le graphique correspondent à des variables aléatoires ; les arrêtes représentent les dépendances statistiques (entre ces variables) qui entraînent le calcul des probabilités communes, conditionnelles et/ou marginales d'intérêt (Airoldi 2007). L'Allocution Latente de Dirichlet a été développée sur la base des réseaux bayésiens hiérarchisés (Blei, Ng, et Jordan 2003). La première application fut réalisée dans la fouille textuelle : il s'agissait de retrouver dans un document les thèmes abordés et les mots qui y sont associés (Blei, Ng, et Jordan 2003). D'autres approches ont été proposées pour l'apprentissage statistique basé sur l'estimation de densités de probabilités (*The Elements of Statistical Learning*» Hastie,

Tibshirani, et Friedman 2009). La méthode d'estimation par noyau (Kernel) est une de ces approches, introduite en 1962 par Emanuel Parzen (Parzen 1962). Sur la base de cette méthode l'algorithme du "K plus proche voisin" (*KNN : K-nearest neighbor*) a été développé pour résoudre des problématiques de classifications notamment pour la classification des documents textuels (Bijalwan et al. 2014). Les méthodes à base d'arbres de décision constituent d'autres approches statistiques utilisées pour la fouille des données. Nous présenterons dans ce qui suit, une application des modèles des forêts aléatoires de décisions qui en est une variante.

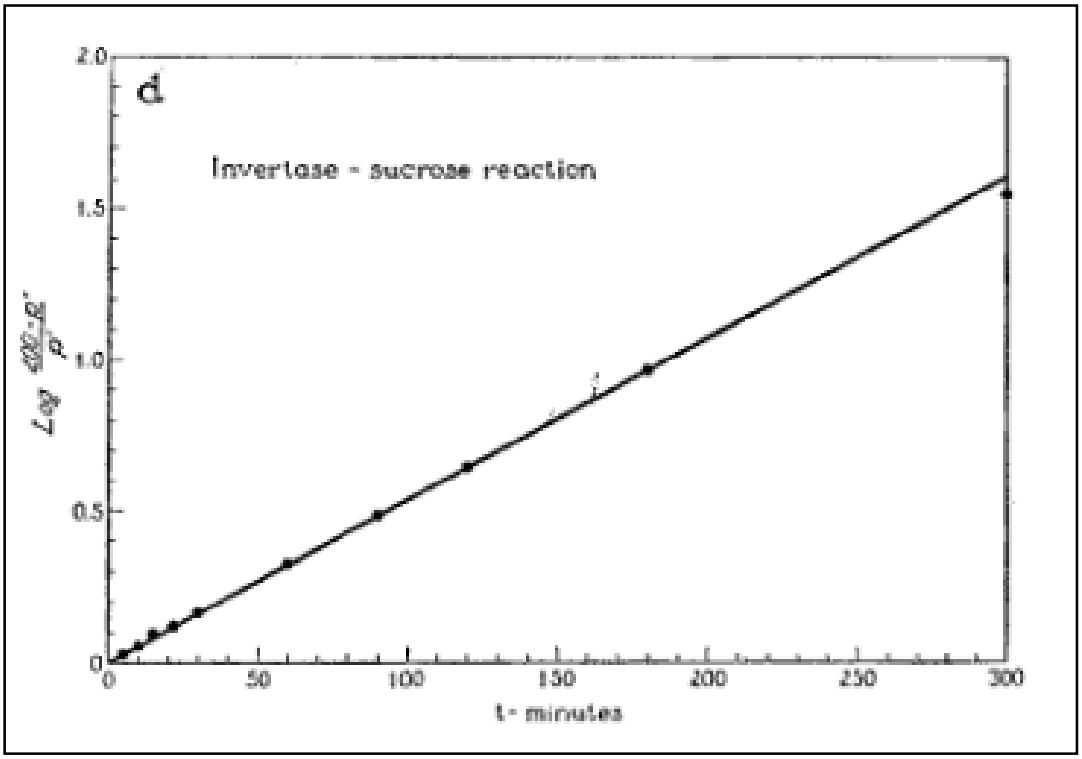


Figure 11. Application du modèle LOGIT à l'hydrolyse du *sucrose* par l'*invertase* en fonction du temps. p' traduit ici le taux de sucrose dans le milieu. Source (Berkson 1951).

1.2.2. L'apprentissage automatique

Selon Mitchell (Mitchell 2006), on pourrait définir l'apprentissage automatique comme la discipline informatique qui permet de répondre à la question suivante : « Comment construire des systèmes informatiques qui s'améliorent automatiquement avec l'expérience, et quelles sont les lois fondamentales qui régissent tous les processus d'apprentissage ? ». L'apprentissage automatique est une discipline à l'interface de l'informatique et des statistiques qui permet la production de systèmes qui améliorent leurs performances pour l'exécution de tâches particulières par l'expérience (parce qu'ils sont capables d'apprendre à partir des données)

(Iniesta et al. 2016). On peut résumer les objectifs de l'apprentissage automatique par les cinq fonctions suivantes (Alpaydin 2010) :

- La classification : l'apprenant doit fournir un modèle qui permet de consigner des données en entrée à des groupes connus de manière supervisée. Une problématique très connue est la détection des spams dans les mails (Kishore Kumar et al. 2012).
- Le regroupement (*clustering*) : le but du regroupement est d'assigner des données en entrée à des groupes à la différence de la classification le nombre des groupes à la sortie n'est pas toujours connu (Hassani et al. 2016) et on se base sur une fonction de similarité. Cette propriété fait des approches de regroupement, des approches non supervisées par excellence.
- L'estimation de densité : il s'agit ici de trouver la distribution d'une population à partir des données d'un échantillon (Donoho et al. 1996).
- La réduction de dimension : c'est processus de réduction des variables à considérer en entrée (Roweis and Saul 2000). Cette technique est utilisée comme prétraitement pour la production des modèles pour réduire le nombre de paramètres, ou bien pour la visualisation.
- Certains auteurs considèrent aussi la régression comme une technique d'apprentissage automatique (Iniesta et al. 2016).

On peut classer les algorithmes d'apprentissage en trois grandes catégories (Russell et al. 1995, Love 2002, Mohri et al. 2012) :

- Les méthodes supervisées : l'apprenant reçoit des données marquées pour l'entraînement et fait des prédictions et des regroupements pour tous les points invisibles. C'est le scénario le plus courant. Il est associé à la classification, à la régression et aux problèmes de classement. Le problème de détection de spam est une instance de classification supervisé.
- Les méthodes non supervisées : l'apprenant reçoit des données d'entraînement non marquées et fait des prédictions pour tous les points invisibles. Étant donné qu'en général, aucun exemple marqué n'est disponible, il peut être difficile d'évaluer quantitativement la performance d'un apprenant. Le regroupement par les k-moyenne et la réduction de dimensionnalité tel que l'analyse en composantes principales en sont des exemples.
- Les méthodes de renforcements : la phase d'apprentissage et la phase de test sont inter-mixées. Pour collecter des informations, l'apprenant interagit avec son environnement et l'affecte. L'algorithme reçoit un retour de l'environnement qui le guide dans son apprentissage.

D'autre catégories ont été décrites dans la littérature, comme les méthodes semi-supervisées (Mohri et al. 2012). Le corpus d'apprentissage est constitué de données marquées et de données non marquées pour faire ces prédictions. On peut recourir à des méthodes probabilistes pour prédire les étiquettes des données non marquées. On parle alors d'inférence transitive.

On pourrait résumer la méthodologie globale en cinq étapes comme présenté dans la Figure 12. Mohri et coll. (Mohri et al. 2012) considèrent que la généralisation est une étape à elle seule. Il s'agit de généraliser le modèle produit comme système d'aide à la décision. De notre côté, comme on va le voir plus tard, la généralisation consiste à généraliser les nouvelles connaissances et/ou prédiction du système. En effet, le modèle généré a été optimisé sur des données produites dans un contexte bien particulier. Si on change les données, ceci pourrait changer les performances du modèle.

Une grande communauté d'informaticiens s'est constituée pour la production d'algorithmes de fouille des données en combinant des approches informatiques et statistiques et il est difficile de tous les énumérer. Dans ce qui suit, je vais présenter quelques applications de la vie courante tout en expliquant les approches techniques et informationnelles utilisées.

1.2.3. Analyse linguistique

C'est une discipline de l'intelligence artificielle qui a été initié par la recherche militaire depuis les années 50, au début pour le décryptage des messages codés (Auroux 1998). Comme nous l'avons vue précédemment c'est une discipline indissociable de la fouille des données textuelles. Elle se base sur l'analyse lexicale, syntaxique et sémantique des textes.

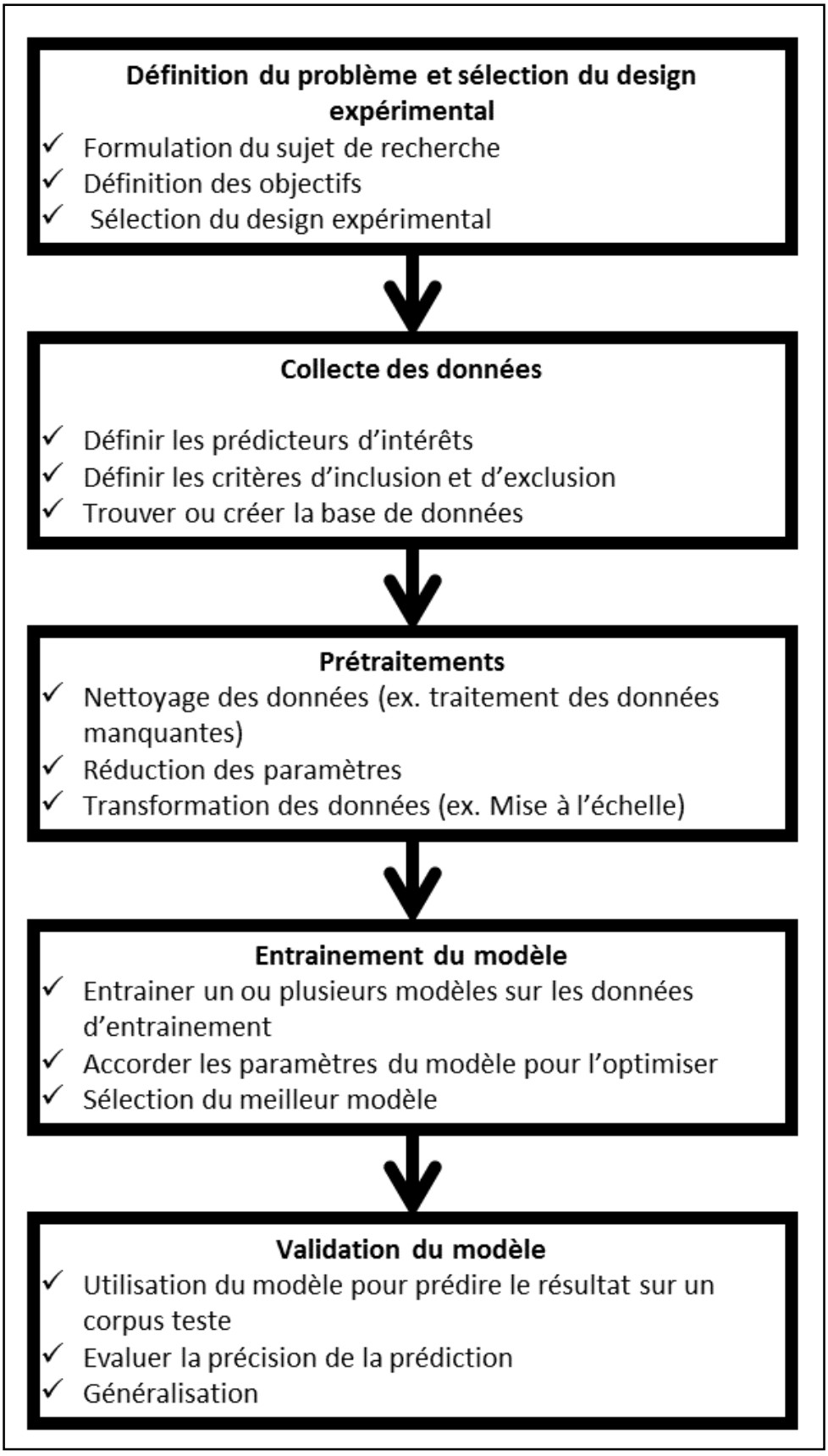


Figure 12. Les étapes principales du processus d'apprentissage en apprentissage automatique. Source (Mohri et al. 2012).

On peut résumer les champs d'applications comme suit (Tan 1999) :

- La recherche d'informations : dans le contexte de l'analyse textuelle, il s'agit d'une étape préliminaire pour la récupération d'un corpus textuel sur le web ou un autre système de fichier.
- Le traitement automatique du langage naturel : l'ensemble des méthodologies informatiques pour l'analyse linguistique telles que l'étiquetage morphosyntaxique
- Reconnaissance d'entités nommées : l'utilisation de nomenclatures ou de techniques statistiques pour identifier des objets textuels (mots ou groupe de mots) catégorisables tels que les personnes, les organisations, les noms de lieux, les symboles... La désambiguïsation, par l'utilisation d'indices contextuels, peut être nécessaire pour décider où, par exemple, « Paris » peut se référer à une personne ou le nom d'une ville.
- La reconnaissance de motifs : caractéristiques telles que les numéros de téléphone, adresses e-mail, les quantités (avec unités) peuvent être discernées par expression régulière ou autre motif correspond.
- Identification des relations : identification des interactions entre objet dans les textes (en biologie, par exemple, une des applications les plus étudiées en fouille des textes est l'identification des interactions entres protéines (Hou et al. 2017)).
- L'analyse du sentiment consiste à extraire diverses formes d'information sur l'attitude : le sentiment, l'opinion, l'humeur et l'émotion.
- La classification textuelle : en fonction de l'objectif et des critères de classification, des techniques de regroupement des modèles d'apprentissage peuvent être appliqués.

L'analyse linguistique et la fouille des textes en général trouvent des applications dans divers domaines allant des prédictions des résultats politiques (Rincy and Varghese 2015) et le Politoscope développé par l'Institut des Systèmes Complexes de Paris Îles-de-France et à la médecine (Gonzalez et al. 2016).

2. Applications des approches des sciences de l'information et des sciences de données

Dans cette partie, j'expose deux applications économiques où les nouvelles approches de fouille des données ont révolutionné les pratiques en usage.

2.1. Prédiction de la pauvreté par analyse des images satellitaires

Cette étude a été présentée dans la revue *Science* par une équipe de l'université de Stanford (Jean et al. 2016).

2.1.1. Problématiques liées à l'analyse des images satellitaires pour la prédiction des données économiques

- **Les images satellitaires pourquoi ?**

La mesure de l'indice de pauvreté est une mesure statistique qui se base sur des enquêtes. Selon un rapport de la Banque Mondiale entre 2000 et 2011 sur 59 états, 39 ont mené moins d'une enquête pour la mesure de cet indice et 14 pays n'ont mené aucune enquête[22]. Pour contourner ce manque d'information, d'autres techniques ont été développées pour la collecte d'information à partir des réseaux sociaux et des supports mobiles. Henderson et coll. (Henderson et al. 2012) ont présenté une étude pour prédire l'activité économique des pays à partir des images satellitaires de luminosité durant la nuit.

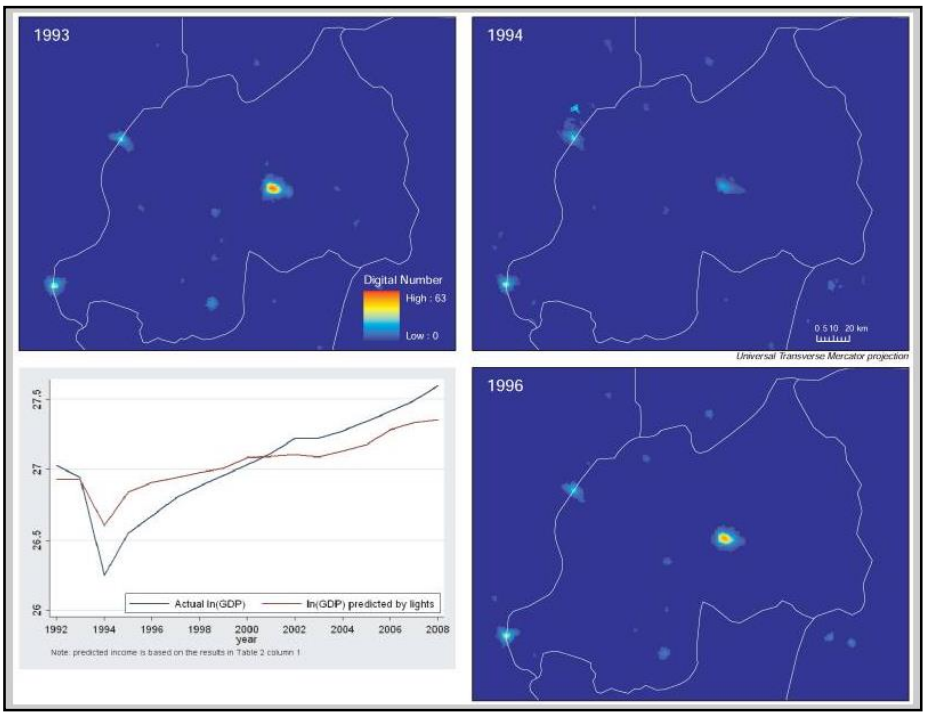


Figure 13. Utilisation des images de luminosité pour la prédiction du PIB du Rwanda. La courbe représente l'évolution réelle du PIB (en bleu) et la prédiction du modèle (en rouge). On note ici que le modèle arrive même à prédire la crise de 1994, où la guerre civile a induit une baisse du PIB (Henderson et al. 2012).

[22] http://iresearch.worldbank.org/povcalnet/home.aspx

Le modèle présenté par Henderson est puissant pour la prédiction de l'activité économique mais a des limites quant à l'analyse de la pauvreté. Jean et coll. (Jean et al. 2016) proposent un modèle basée sur l'extraction des données socio-économiques à partir d'images satellitaires à haute résolution du jour.

- **Problématique liée aux données**

Parmi les données indispensables au calcul de l'indice de la pauvreté, on compte les enquêtes démographiques et sanitaires, et les données sur les actifs des ménages. En cas d'absence de ces données, l'urbanisme du pays pourrait fournir des données qui pourraient renseigner sur le taux de pauvreté. Cette piste a été explorée dans cette étude. Les données d'urbanisme pourraient être déduites à partir des images satellitaires à haute résolution de jour, contrairement à la méthode de Henderson et coll. (Henderson et al. 2012). Malheureusement, de telles données sont très peu structurées, et présente des défauts de résolution rendant l'extraction de l'information difficile, même avec une analyse manuelle intensive. En effet, sur ces images, on collecte des informations telles que les routes, les habitations, les zones urbaines et les zones rurales (Figure 14).

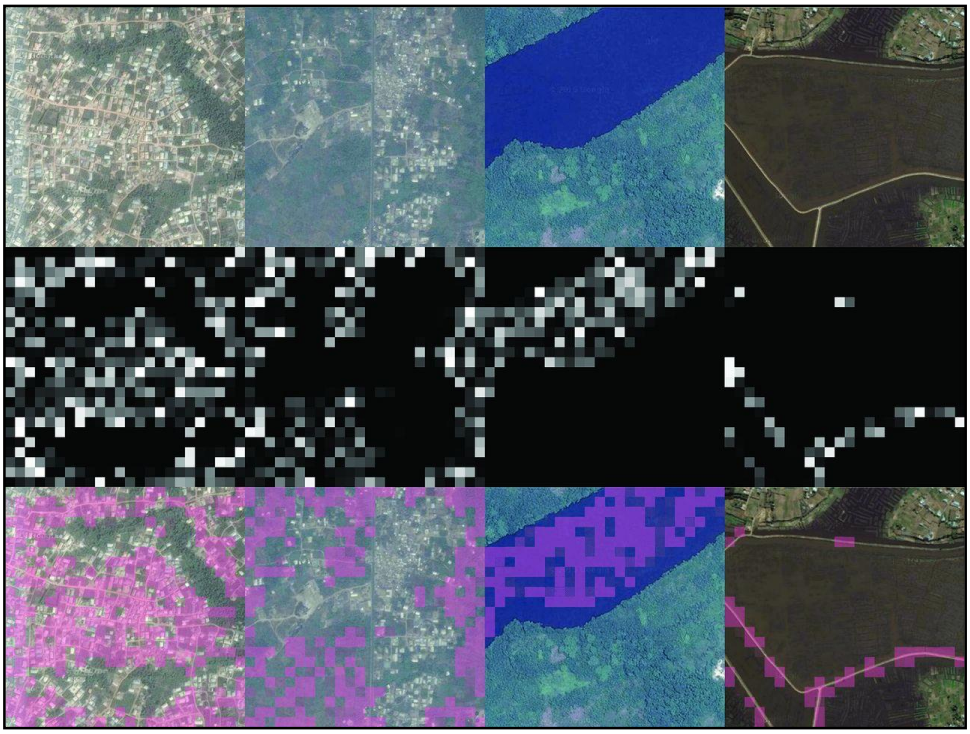


Figure 14. Images satellites traitées. Par colonne : Quatre différents filtres qui identifient, de gauche à droite, des caractéristiques correspondant aux zones urbaines, aux zones non urbaines, à l'eau et aux routes dans le modèle. Chaque filtre "met en surbrillance" les parties de l'image, affichées en rose. Par ligne : des images satellitaires de jour provenant de Google Static Maps, des cartes d'activation des filtres et des cartes d'activation sur les images d'origine (Jean et al. 2016).

2.1.2. Approche utilisée pour l'analyse des images satellitaires pour l'étude de la pauvreté

Pour l'analyse des données, les auteurs proposent une approche basée sur l'apprentissage par transfert (*Transfer Learning*) (Figure 15).

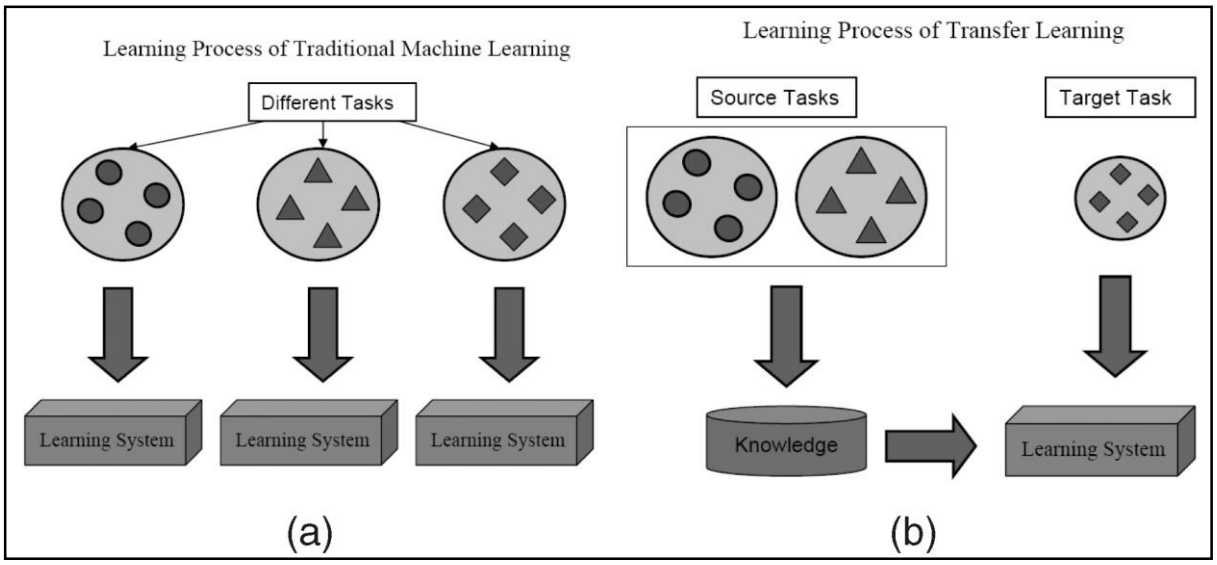


Figure 15. Différence entre les approches traditionnelles d'apprentissage (a) et l'apprentissage par transfert (b). Dans le premiers exemple (A) si l'exercice est de de reconnaitre les formes géométriques suivantes cercle, triangle et losange, l'algorithme doit s'entrainer sur des cercles triangles et losanges. Dans l'apprentissage par transfert si l'algorithme s'entraine seulement à reconnaitre des cercles et des triangles il saura aussi à reconnaitre des losanges. Sources : (Pan and Yang 2010).

- **Apprentissage par transfert**

Selon Pan and Yang (Pan and Yang 2010), contrairement aux techniques d'apprentissage supervisées, dans l' « apprentissage par transfert », les données d'apprentissage peuvent être totalement différentes. On assimile cela à quelqu'un qui apprend à jouer du clavier et saurait jouer aussi du piano.

- **L'approche proposée par Jean et coll.**

La première étape consiste à entraîner un modèle de réseaux de neurones convolutionnel (CNN : *Convolutionnal Neural Network*) pour l'analyse des images satellitaires. Un réseau de neurones artificiel est un algorithme bio-inspiré très utilisé en apprentissage automatique. C'est un réseau fortement connecté de processeurs élémentaires qui fonctionnent parallèlement. Chaque processeur calcule une sortie sur la base des informations qu'il reçoit. Le CNN est un réseau de neurones avec une architecture particulière inspirée par le cortex visuel animal.

La deuxième étape consiste à réduire le nombre de dimensions des vecteurs de sortie du modèle CNN ; les auteurs utilisent l'ACP (L'Analyse en Composante Principales) pour diminuer les temps de traitement.

Pour l'estimation de la pauvreté, les données collectées sont introduites dans un modèl basé sur la régression linéaire. Pour l'apprentissage de ce modèle, les auteurs ont utilisé des données issues des enquêtes démographiques et sanitaires et celles issues d'enquêtes de la Banque Mondiale. Ces données subissent un prétraitement ; elles sont regroupées. La prédiction se fait sur les valeurs de pauvreté moyenne par regroupement.

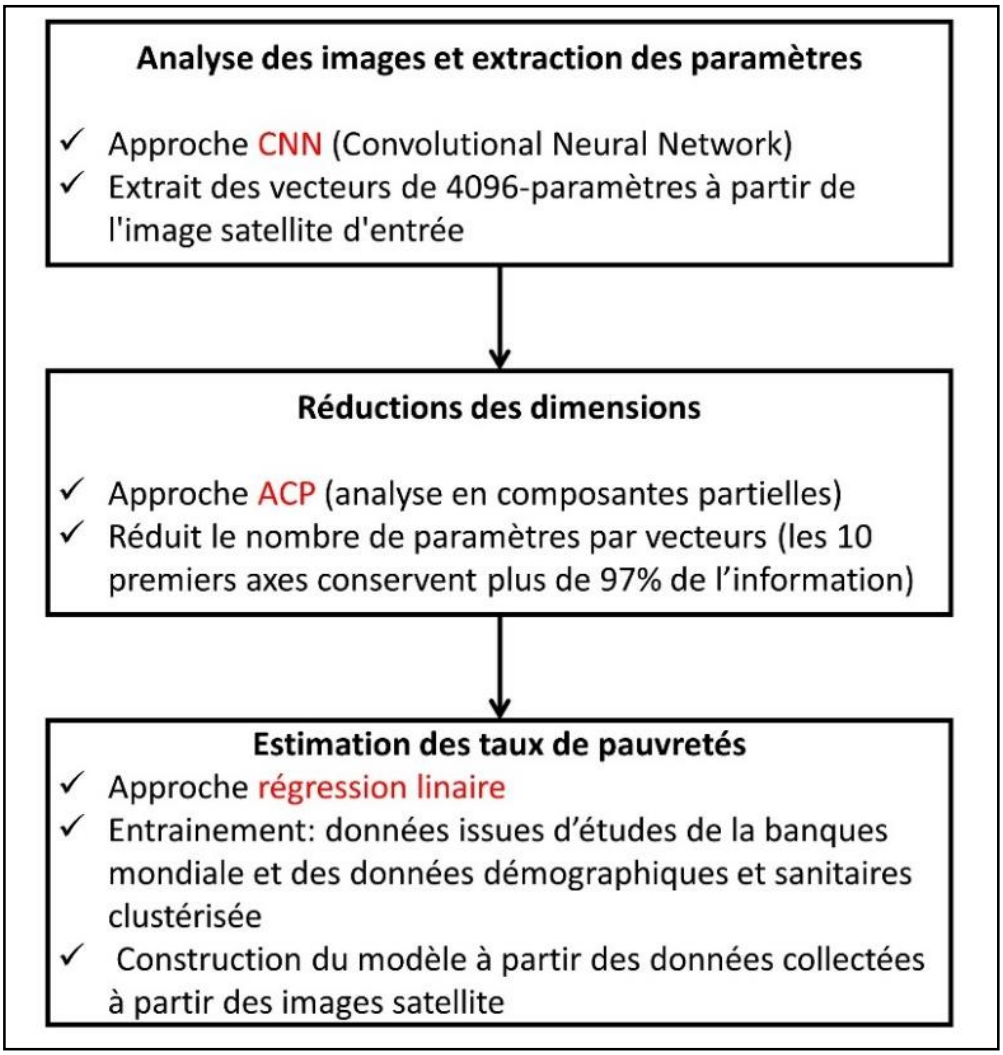


Figure 16. Principales étapes de l'approche développée par Jean et coll. pour la prédiction de la pauvreté à partir des images satellitaires. L'approche commence par s'entrainer à extraire les paramètres à partir d'image satellitaire (routes, maisons, village…). L'Analyse en Composantes Principales permet de réduire le nombre de paramètres. Ces paramètres sont combinés dans un modèle de régression linéaire pour l'estimation du taux de pauvreté, les données de la banque mondiale ont servi de données d'apprentissage. Sources : (Jean et al. 2016).

2.1.3. Résultats obtenus pour la prédiction de la pauvreté à partir d'images satellitaires

Les résultats présentés par Jean et coll. (Jean et al. 2016) montrent que l'utilisation des images satellitaires de jour peut donner des prédictions sur la répartition spatiale du bien-être économique plus précises que les images de nuit. Néanmoins, les prédictions du modèle montrent une variabilité entre les données prédites et les données observées de 37% à 55% si on entraîne le modèle pays par pays. Si on entraîne le modèle sur des données groupées pour tous les pays cette variabilité augmente à 44%-59%. Ceci pointe une des faiblesses de l'apprentissage automatique : la généralisation d'un modèle entraîné sur les données d'un pays pour la prédiction des données d'un autres pays est difficile. En effet la performance des modèles dépend beaucoup des données d'entraînement. Le modèle entrainé pays par pays est plus sensible que le modèle entraîné par le groupement des données.

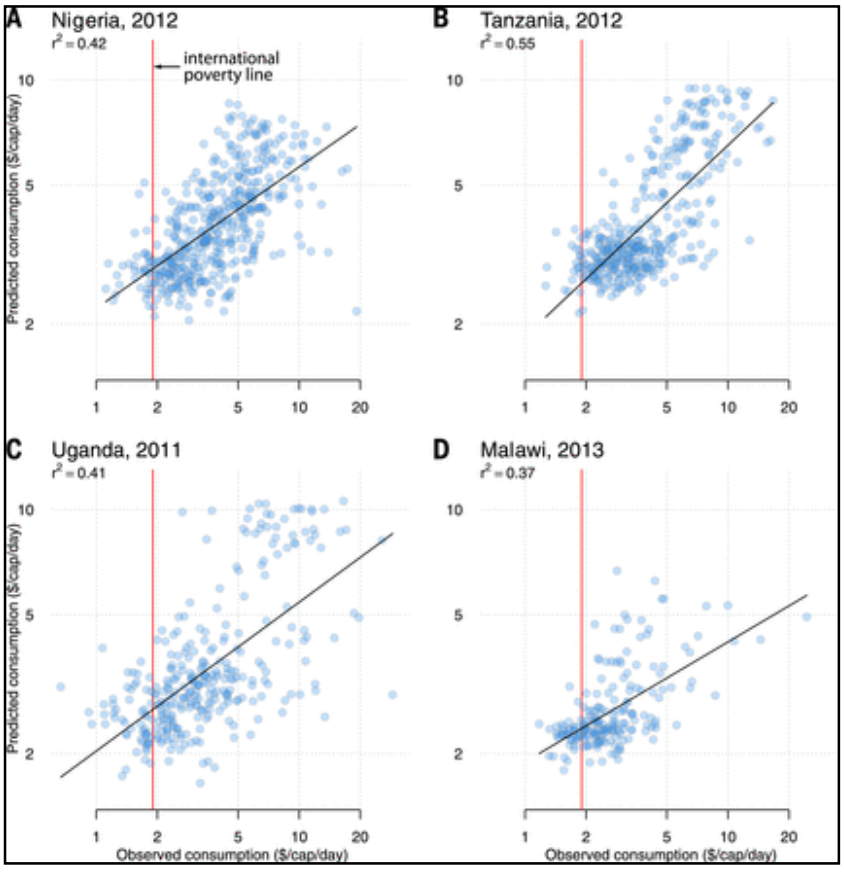


Figure 17. Consommation prédite par le modèle d'apprentissage par transfert (l'axe des y) par rapport à la consommation observée par cluster pour 4 pays africains. Les résultats sont montrés pour 4 pays (A) le Nigéria, (B) la Tanzanie, (C) l'Uganda et (D) le Malawi. Les prédictions et les valeurs de r^2 rapportées pour chaque panel après 5 tours de validation croisée. La ligne noire représente le meilleur ajustement, et la ligne rouge est le seuil de pauvreté international de 1,90 $/personne/jour. Les deux axes sont représentés en échelle logarithmique. Source : (Jean et al. 2016).

2.2. Prédiction des récoltes

Cette étude a été présentée par une équipe de l'université de Washington dans la revue *Plos One* en 2016 (Jeong et al. 2016).

2.2.1. Problématique liée à la prédiction des récoltes

La capacité de prédire le rendement des cultures en réponse à la variabilité climatique à l'échelle régionale et globale est cruciale pour l'élaboration de politiques agricoles, la prévision et l'analyse des tendances du commerce mondial et l'identification de stratégies d'adaptation efficaces aux changements climatiques. Historiquement, deux approches ont été développées pour cette prédiction les modèles basés sur les procédés de cultures et les modèles basés sur l'estimation statistique. En fonction du modèle des limites peuvent exister :

- Les modèles basés sur les procédés de culture sont très puissants quant à l'estimation des rendements à l'échelle du champ car ils peuvent simuler les réponses physiologiques face aux changements climatiques et aux pratiques agricoles. Ces modèles présentent un problème quand on passe à l'échelle régionale ou nationale (voir Figure 18) de calibrage et de paramétrage vu la taille des données.
- Les modèles statistiques tels que la régression linaire simple et multiple sont des modèles robustes pour la prédiction ; néanmoins, ils ne sont pas très informatifs.

2.2.2. Approche pour la prédiction des récoltes

Jeong et coll. (Jeong et al. 2016) présentent un modèle basée sur l'apprentissage par les forêts d'arbres décisionnelles (*random forest classifier*) et ils l'ont comparé à la régression linaire multiple.

- **Les forêts d'arbres décisionnels**

C'est une méthode de classification ou de régression d'apprentissage supervisé. Le principe est de générer une multitude d'arbre de décision[23] au moment de l'apprentissage ; à la sortie du modèle, on prend la classe qui correspond à la classe majoritaire (Hastie et al. 2013).

Les auteurs proposent de développer un modèle qui permet de combiner une dizaine de paramètres physico-chimiques. Ces paramètres prédictifs contiennent des variables

[23] L'arbre de décision en apprentissage utilise un arbre de décision conventionnel comme modèle prédictif qui permet de faire des observations sur un élément (ces observations sont représentées dans les branches) et de tirer des conclusions sur la valeur cible de l'élément (représentées dans les feuilles).

environnementales incluant le climat, le sol, la photopériode, l'eau et les données de fertilisation (apport en minéraux).

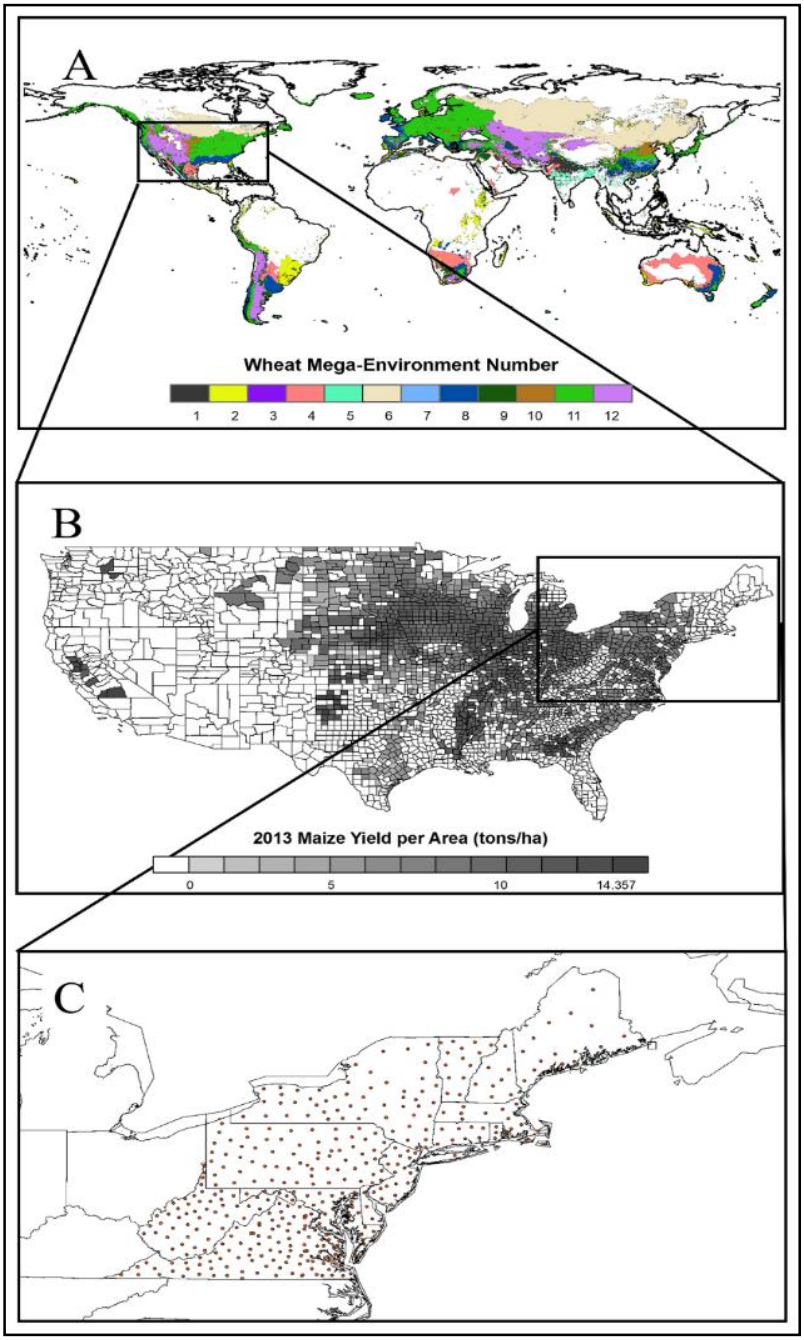


Figure 18. Régions d'étude pour la prédiction des récoltes. (A) Méga-milieux mondiaux du blé, (B) Comtés producteurs de maïs des États-Unis et (C) Région côtière du nord-est des États-Unis. (A) Tous les 12 méga-milieux de blé sont représentés avec des couleurs différentes. (B) Le rendement en grains du maïs par les comtés américains en 2013 est visualisé en utilisant différentes nuances avec des teintes plus foncées représentant des rendements plus élevés. (C) La région côtière du nord-est des États-Unis comprend 433 comtés. Les points rouges indiquent l'emplacement des points de données, où existent des stations météorologiques. Des données de type point ont été utilisées pour cette région. Source : (Jeong et al. 2016).

2.2.3. Résultats obtenus pour la prédiction des récoltes par la méthode des forêts décisionnelles

Table 2. Statistiques générales pour l'évaluation des modèles de forêt décisionnelle et régression linéaire multiple pour la prédiction des récoltes. Source : (Jeong et al. 2016).

Culture	Echelle	Forêt décisionnelle			Régression linéaire multiple		
		EQ	EF	D	EQ	EF	D
Blé	Mondiale	0,32	0,96	0,99	1,32	0,31	0,68
Maïs (grains)	Etats-Unis	1,13	0,76	0,92	1,93	0,30	0,67
Pomme de terre	Côte-est des Etats Unies	2,77	0,75	0,95	5,62	-0,87	0,73
Maïs (ensilage)	Côte-est des Etats Unies	1,90	0,85	0,97	4,54	-0,41	0 ,75

EQ : Erreur quadratique ; elle renseigne sur le taux d'erreur entre les valeurs prédites et les valeurs observées

EF : Coefficient d'efficience ; il varie entre -∞ et 1. Une valeur proche de 1 indique que les valeurs prédites correspondent aux valeurs observées.

D : Indice de Willmot est un indicateur sur la capacité du modèle à prédire. Il varie entre 0 et 1 ; 1 indique une capacité prédictive parfaite.

Les résultats démontrent que le modèle à base d'arbres décisionnels est plus performant que le modèle classique à base de régression linaire. On note aussi que la diminution d'échelle abaisse les performances des deux modèles ce qui confirme aussi notre critique sur la dépendance du modèle du set d'apprentissage et la difficulté de le généraliser malgré les bonnes performances. Pour améliorer les performances il faudrait produire un modèle par échelle.

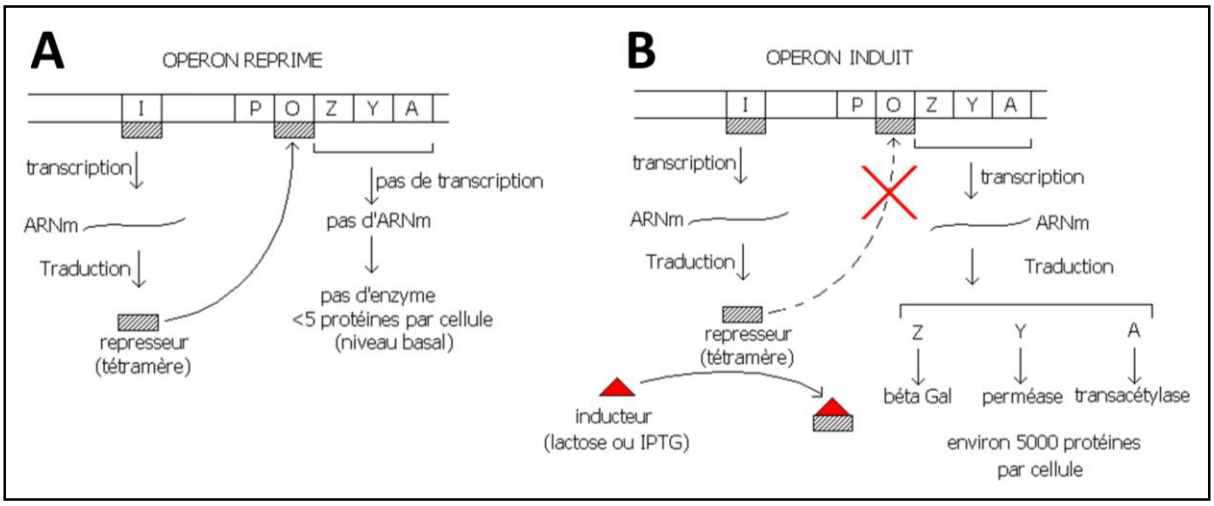
C. La recherche biomédicale et les sciences de l'information

1. Les systèmes biologiques : des systèmes d'information ?

Comme présenté dans la section A2, l'information génétique est l'une des informations les plus anciennes. En effet, les expériences de Mendel (Mendel 1865) ont démontré que les caractères sont transmissibles de génération en génération. Si la structure de l'ADN et son rôle comme support de **l'information génétique** n'a été mis en évidence qu'en 1953 (Watson and Crick 1953), on savait déjà que cette information était déjà structurée en chromosomes (Morgan 1910) et en gènes (Sturtevant 1913). L'élucidation du code génétique (Nirenberg and Leder 1964, Khorana 1968) montre que la succession des nucléotides n'est pas aléatoire. On pourrait

l'interpréter comme un « formalisme structuré », ou bien un « langage » d'où la dénomination **code génétique**. Les expériences de Jacob et Monod confirme cette propriété du code génétique (Jacob and Monod 1961). En effet, ces travaux démontrent que la synthèse protéique se fait par la traduction du code génétique. Cette traduction n'est pas le fruit du hasard mais elle se fait en réponse à un changement de l'environnement (la présence du lactose dans l'expérience induit l'activation de l'opéron Lac). Ce changement de l'environnement induit tout un mécanisme pour la levée de l'inhibition de la transcription de la séquence de l'ADN en ARN, puis la traduction de l'ARN en protéine ; c'est un **mécanisme de régulation** (Figure 19).

Figure 19. Schémas récapitulatifs des mécanismes de la répression et de l'induction de l'opéron lactose chez *Escherichia coli*. A) En absence de lactose dans le milieu le répresseur va empêcher la traduction de l'opéron. B) L'induction de l'opéron se fait par la présence de lactose dans le milieu qui va se fixer sur le répresseur et empêcher son action et il induit ainsi la synthèse de trois protéines (Z : B-galactosidase, Y : perméase et A : transacétylase). Source : http://www.cours-de-biochimie.fr/operons.php.



D'après cette chronologie, nous pouvons conclure que le matériel génétique des cellules biologiques constitue un système d'information avec un système de codage (le code génétique). Ce système d'information possède ces propres mécanismes pour exploiter l'information génétique (ensemble de gènes) en fonction des signaux reçus de la part de son environnement. L'étude de la transcription à haut débit démontre qu'en réponse à un changement de l'environnement, une cellule va activer ou désactiver plusieurs gènes[24] qui vont interagir soit en cascade soit simultanément pour décrire des réseaux de gènes (métaboliques, de transduction, de régulation). Si on extrapole aux systèmes biologiques, un système biologique est un ensemble de molécules (gènes et leurs produits, métabolites…), de cellules, de populations cellulaires et d'organes qui interagissent pour mener une ou plusieurs fonctions

[24] https://www.ncbi.nlm.nih.gov/geo/

biologiques. Les systèmes biologiques peuvent avoir une structure multi-échelle (de la molécule à l'organe), ou non comme dans le cas des réseaux de régulation (Green 2016). Le système immunitaire, dont le rôle est de préserver l'intégrité du corps, contre les agents infectieux, et les dysfonctionnements du soi (Thomas-Vaslin 2015, 2016), est un méta-système multi-échelle composé d'organes de cellules immunitaires impliquant plusieurs mécanismes moléculaires (Richard D. 1992, Litman et al. 2005) (Figure 20).

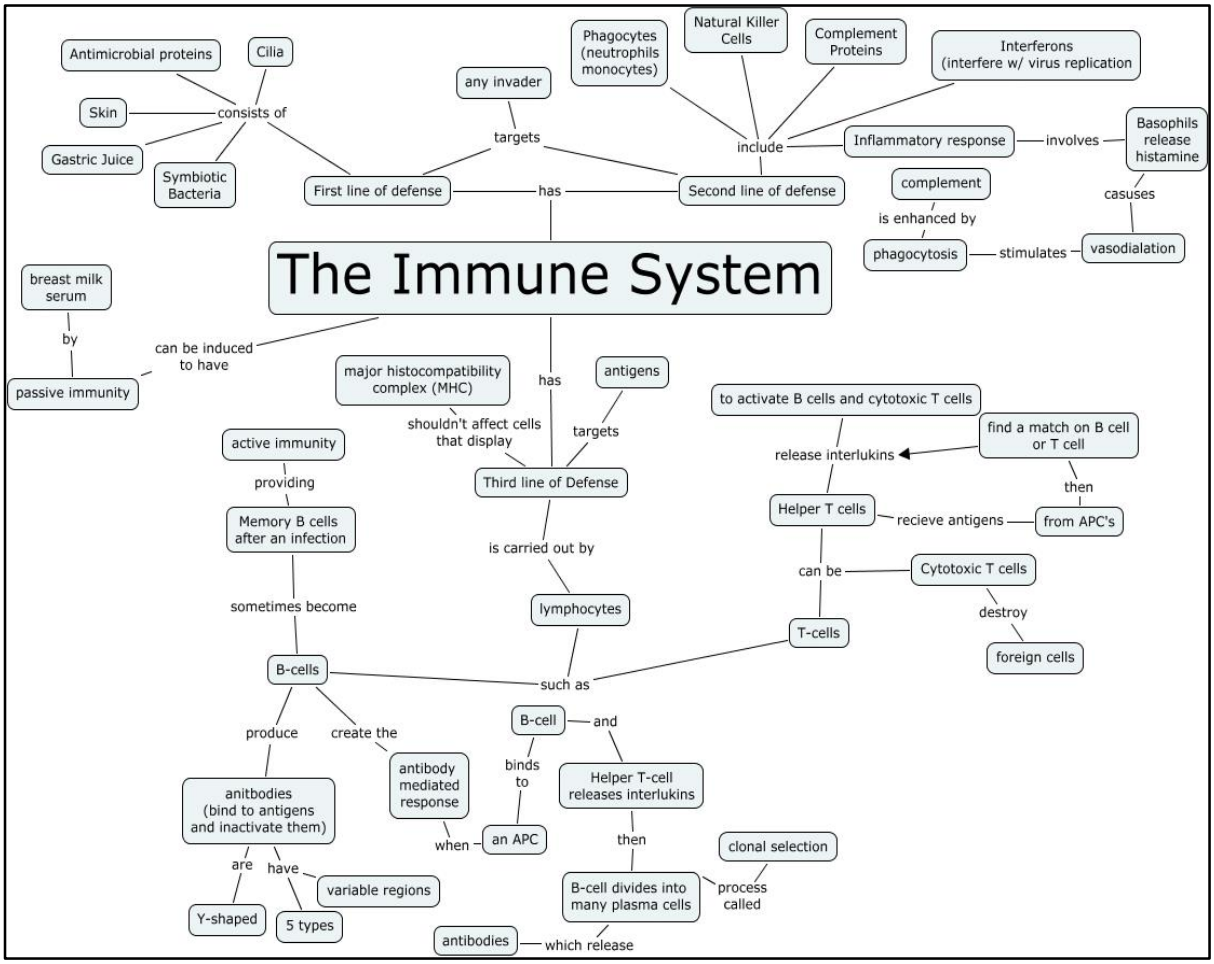


Figure 20. Schéma récapitulatif des populations cellulaires et mécanismes génétiques et moléculaires impliqués dans le système immunitaire. Les organes immunitaires primaires et secondaires ne sont pas représentés ici. C'est une représentation conceptuelle dont la racine est le système immunitaire. Les classes (dans les cadres bleus) sont des espèces cellulaires et moléculaires, des substances ou bien des processus biologiques. Les arêtes représentent les relations entre les différentes classes. Source : http://maaz.ihmc.us/rid=1178472505313_378108100_21726/immune%20system.cmap.

Pour assurer sa fonction, les différentes composantes du système immunitaire doivent communiquer entre elles et avec les autres organes. Le système immunitaire repose sur un système de communication complexe pour la transmission de l'information à travers des canaux biologiques tels que le système sanguin et le système lymphatique, un ensemble de récepteurs moléculaires et cellulaires responsables de la transmission de l'information (Springer 1990),

ainsi qu'une panoplie de médiateurs des voies de transduction du signal. En outre, le système immunitaire est un système biologique à mémoire (Gattinoni et al. 2017) : lors d'un épisode infectieux, par exemple, une partie des cellules effectrices produites est conservé sous la forme de cellules mémoires. Pour conclure, le système immunitaire transmet, interprète, analyse, régit et conserve l'information. Par analogie à l'information électronique, on peut parler d'un système d'information biologique.

Pour l'étude de ces systèmes biologiques, plusieurs techniques ont été développées pour la compréhension, l'analyse et l'interprétation de l'information émise par ces systèmes biologiques. La biologie des systèmes computationnelle est une discipline qui utilise la puissance de la modélisation mathématique, statistique et informatique pour mimer le comportement des systèmes biologiques. Pour ce faire, cette technique se base sur l'exploitation et l'analyse des résultats des expérimentations biologiques (Kitano 2002). Récemment, Bersini et coll. (Bersini et al. 2012) ont développé une approche de modélisation de la dynamique des populations lymphocytaires en utilisant des diagrammes d'états. Le but de cette démarche est de dépasser les limitations liées aux difficultés des approches informatiques et mathématiques pour modéliser des systèmes aussi complexes que le système immunitaire. Par exemple pour la modélisation de la dynamique des Lymphocytes T durant l'homéostasie Thomas-Vaslin et Coll. utilisent un modèle mathématique à base d'équations différentielles ordinaires à six variables et dix-sept paramètres (Thomas-Vaslin et al. 2008). L'élucidation et la compréhension de tels systèmes d'équations nécessitent des compétences poussées en mathématiques. Un autre modèle de la dynamique de différentiation des lymphocytes T dans le thymus a été produit par Souza-e-Silva et coll. (Souza-e-Silva et al. 2009). C'est un modèle informatique utilisant un automate cellulaire, qui présente l'avantage d'intégrer l'effet du microenvironnement thymique (tel que l'effet des chimiokines) par rapport au modèle de Thomas-Vaslin et coll. Néanmoins, le modèle se présente sous la forme d'un code informatique de plus de six cents lignes sous Fortran[25]. Ces modèles restent donc peu accessibles pour des biologistes, qui ne sont pas forcément experts en informatique. Pour comprendre l'intérêt de l'approche de modélisation graphique présenté par Bersini et coll., je présente un exemple dans la Figure 21. Dans cet exemple un diagramme de transition a été produit pour la modélisation de la différenciation thymique en partant du modèle conceptuel de Thomas-Vaslin et coll. (Thomas-Vaslin et al.

[25] https://gcc.gnu.org/fortran/

2008). L'utilisation des équations différentielles ordinaires a nécessité l'utilisation de trente équations pour la description de tous les processus (Thomas-Vaslin et al. 2013).

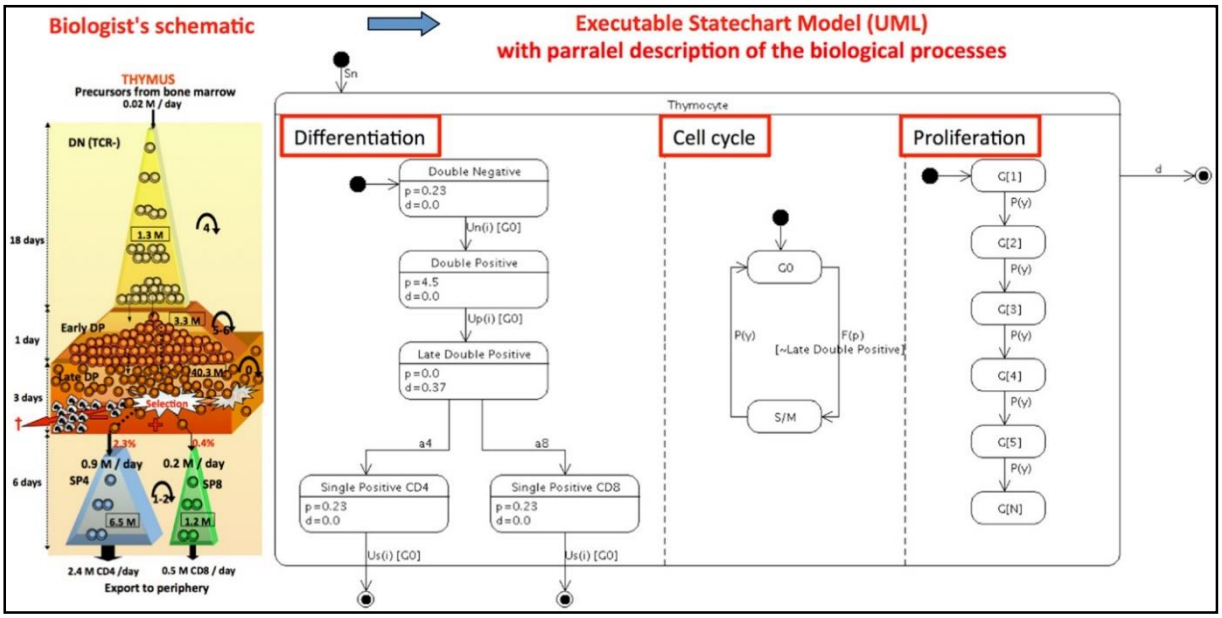


Figure 21. Factorisation du modèle de différenciation des thymocytes. **C'est le modèle proposé par Thomas-Vaslin et coll. en 2012 sous la forme d'équations différentielles dans un langage visuel exécutable par ordinateur (diagramme de transition proposé par Bersini et coll.). A gauche, le modèle conceptuel proposé par les biologistes de la dynamique des thymocytes initialement proposé est transformé en un diagramme de transition UML, à droite. Ce diagramme décrit l'évolution des populations de cellules dans le thymus, représentées par leur état de DN (*Double Negative*) à SP (*Simple positive*). Les entrées du modèle sont représentées par des cercles pleins. Les sorties sont représentées par les doubles cercles. Les flèches représentent le passage d'un état à un autre. Les processus biologiques sont soulignés avec les boîtes rouges et se produisent d'une manière parallèle. Les taux de prolifération cellulaire sont notés « p » et de mort cellulaire « d ». Source : (Thomas-Vaslin et al. 2013).**

L'approche de Bersini et coll. est plus accessible pour les non spécialistes de l'informatique et des mathématiques. Néanmoins, cette approche se base sur une description fine des phénomènes biologiques sous-jacents et une analyse de résultats expérimentaux pour le paramétrage et la validation des modèles. Dans notre exemple (Figure 21), trois processus biologiques sont représenté par des cadres rouges mais des processus de mort sont mis en jeux et sont représentés de façon implicite. L'analyse de la littérature et des bases de données biologiques pourrait constituer une source fiable de données expérimentales et pour la description des processus. D'où l'importance de l'étape d'analyse et d'interprétation des résultats transcriptomiques dont on dispose au laboratoire mais aussi qu'il faudra lier aux observations effectuées aux différents niveaux d'échelles du système immunitaires (population cellulaire, organe et organisme).

2. Les systèmes biologiques : générateurs d'information

Il est difficile de quantifier l'information scientifique générée par la biologie. Néanmoins, la biologie en tant que science a profité du développement des technologies de l'information. MEDLINE®[26], le portail bibliographique de la bibliothèque nationale de médecine des Etats-Unis, base de données bibliographique de référence pour le domaine biomédical, nous renseigne sur l'évolution de la production scientifique biologique à partir du 20ème siècle (Figure 22). Un autre indicateur de l'abondance de l'offre d'informations en biologie est la multiplication des bases de données spécialisées. Le site du NCBI[27], portail du Centre National pour l'Information Biotechnologique, compte plus de 70 bases de données de références. Le portail du laboratoire européen de bio-informatique[28], quant à lui, en compte plus d'une cinquantaine. Cette explosion de l'offre informationnelle est la conséquence du développement des technologies expérimentales utilisées en biologie et en médecine comme les techniques de séquençage massif.

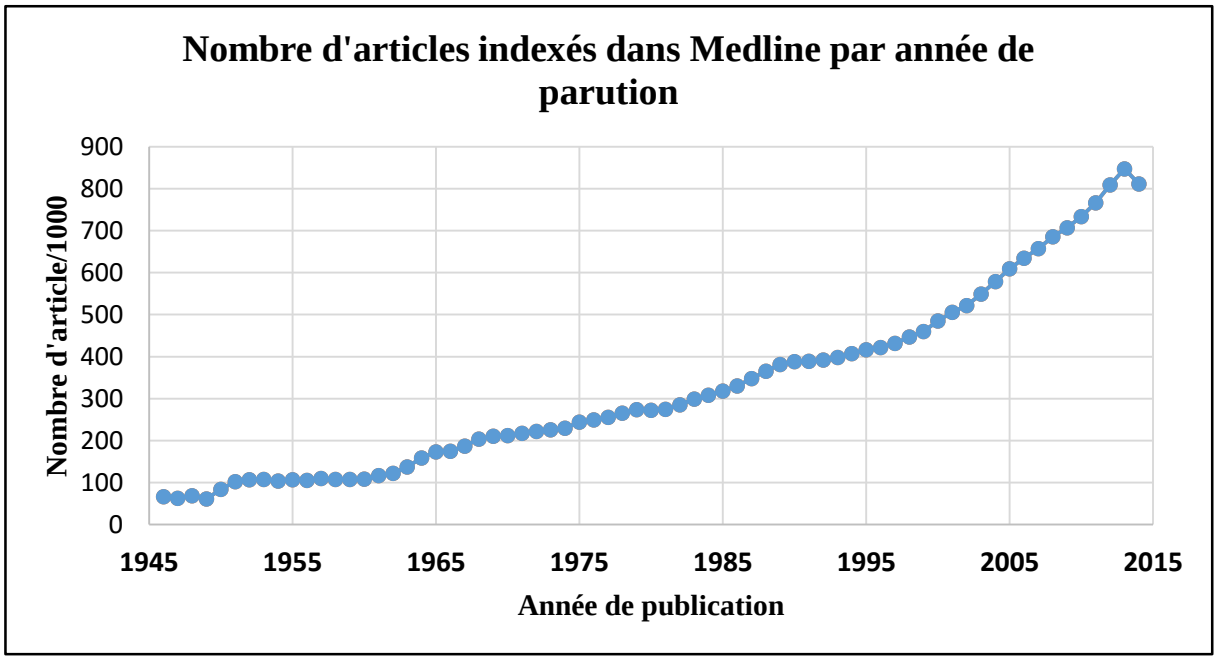


Figure 22. Courbe représentative du nombre de publications indexées dans MEDLINE par année de parution. Cette statistique a été faite courant 2015. Ici on présente les données de 1947 à 2013. Source : https://www.nlm.nih.gov/bsd/medline_cit_counts_yr_pub.html

[26] https://www.nlm.nih.gov/bsd/pmresources.html

[27] https://www.ncbi.nlm.nih.gov/guide/

[28] http://www.ebi.ac.uk/services/all

1.1. Evolution des technologies de séquençage

En 1990, commençait le projet de séquençage du génome humain. Le but étant de produire la première séquence complète du génome humain composé de quelque trois milliards de paires de base. Ce projet a mobilisé une vingtaine d'équipes mondiales (Lander et al. 2001). Les premières techniques de séquençage ont été mises au point dans les années 1970 (Maxam and Gilbert 1977, Sanger et al. 1977). Ces techniques révolutionnaires présentaient plusieurs limites en raison de la lenteur des protocoles, du coût et de la limite de longueur des séquences produites. Malgré les efforts qui ont été entrepris pour l'automatisation, des techniques de séquençage (Meldrum 2000a, 2000b) et l'amélioration des protocoles pour l'optimisation de la capacité de séquençage qui a atteint 100000 séquences en 12 heures (Lander et al. 2001), la séquence du premier chromosome humain n'a été publié qu'en 1999 (Dunham et al. 1999) et le premier génome humain complet a été publié courant 2001 (Lander et al. 2001). Une autre séquence du génome humain a été publiée dans la foulée par l'entreprise Celera Genomics (Venter et al. 2001). Les auteurs de cette publication assurent avoir réussi à séquencer plus de 14 milliards de paires de bases en neuf mois. Ces avancées annoncent alors l'entrée dans l'aire de la génomique et la production des données biologiques de masse (haut débit). Si les avancées du projet génome humain ont contribué à l'amélioration des techniques, les coûts aussi ont été abaissés grâce à ce projet (Figure 23). Néanmoins, la démocratisation de la génomique a été effective avec l'apparition des nouvelles technologies de séquençages à haut débit. La commercialisation de plates-formes de séquençage, comme le séquenceur 454 de l'entreprise Roche, a permis le séquençage du 1er million de paires de base du génome néandertalien (Green et al. 2006). Son utilisation a été étendue à d'autres applications, notamment pour l'analyse qualitative et quantitative de l'ARN messager telle que l'analyse du répertoire des lymphocytes T (Fang et al. 2014, Thapa et al. 2015). Les plates-formes Illumina sont très utilisées aussi pour le séquençage à haut débit des acides nucléiques (Shendure and Ji 2008). Depuis 2014 l'entreprise illumina a annoncé la mise sur le marché nouvelles plateformes de séquençage pour réduire le coût du séquençage et atteindre l'objectif du génome humain à 1000$. Les nouvelles plateformes permettent de réduire le processus de prétraitement (ce qui permet de réduire les coûts du séquençage), et sont capables de réaliser de nouvelles analyses tel que le séquençage d'une cellule unique c'est le cas des plateformes Ion Torrent et Illumina (Buermans and den Dunnen 2014). D'autres plateformes comme la plateforme PacBio offre l'avantage d'être moins gourmande en réactifs et en ADN cible, elle génère aussi des fragments plus longs nécessitant

un temps de traitement informatique plus court, mais le taux d'erreurs est un peu plus important que les autres plateformes (Buermans and den Dunnen 2014).

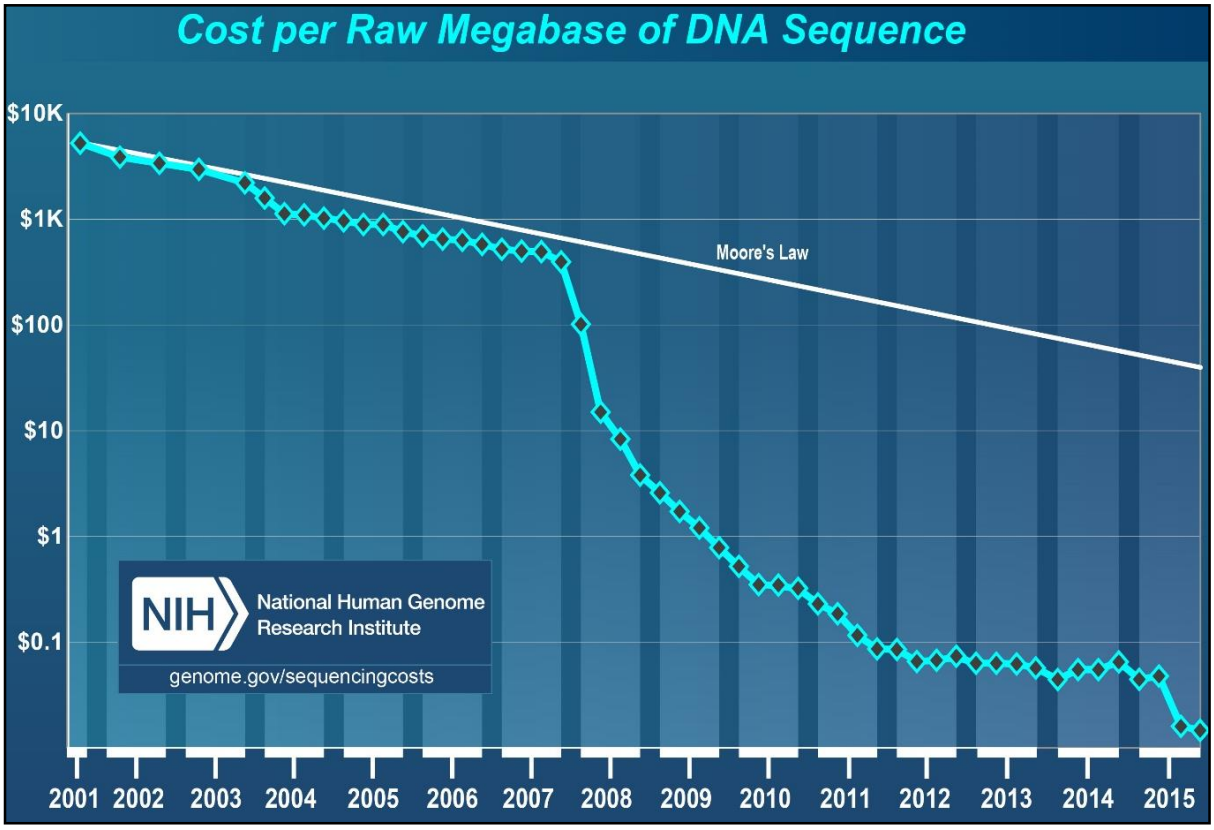


Figure 23. Evolution du coût du séquençage par million de paires de bases entre 2001 et 2015. La courbe bleue représente le coût réel et la courbe blanche représente le coût prédit selon la loi de Moore. Source : https://www.genome.gov/sequencingcostsdata/

L'avènement de ces technologies de séquençage et de quantification des acides nucléiques à haut débit a permis la diminution des coûts du séquençage mais aussi l'afflux de données massives qu'il va falloir analyser et interpréter. Les techniques bioinformatiques telles que l'alignement de séquences permettent à partir des données de séquences de prédire les régions codantes ou non, de voir le degré de conservation d'une séquence. Pour l'étude de l'implication des observations expérimentales (les mutations, la sur-expression et la sous-expression des gènes) dans les processus biologiques, il faut interpréter ces résultats, le but étant de voir l'effet sur les processus biologiques. Classiquement, le processus d'interprétation se base sur les connaissances des expérimentateurs. La taille des données générées par les nouvelles technologies expérimentales pose un nouveau challenge pour l'automatisation des processus d'analyses et d'interprétations. Dans ce sens, la découverte de connaissances pour

des tâches telles que la prédiction des exons-introns et les structures des protéines à partir de la séquence et l'inférence des réseaux de régulation génique à partir des profils d'expressions est un nouveau paradigme de la bioinformatique (Kitano 2002). Ces méthodes utilisent généralement des méthodes inspirées des sciences de l'information et des sciences de données, telles que les prédictions basées sur des heuristiques, sur des discriminateurs statistiques et sur d'autres algorithmes basés sur la linguistique. On pourrait résumer ces étapes au processus d'annotation (Kawai et al. 2001, Stein 2001). L'automatisation du processus de l'étape d'interprétation passe par le criblage automatique des bases de données biologiques et de la littérature. Cette problématique est commune à toutes les technologies expérimentales à haut débit telles que les puces génomiques et transcriptomiques. Le développement de tels outils ne peut se faire sans la standardisation des terminologies biologiques.

1.2.Efforts de standardisation du langage biologique

Le problème de standardisation des terminologies en biologie est ancien. Il a été évoqué précédemment pour les noms de gènes[29]. Shows et coll. (Shows et al. 1987) ont édité les premières lignes directrices pour avoir une terminologie standardisée et unifiée pour les noms de gènes et de protéines humaines, et leur abréviation, mais aussi pour la description des phénotypes et des maladies héréditaires. Pour maintenir cette nomenclature, éviter les redondances (le principe est « un nom, un gène ») et maintenir une base de noms uniques, un comité a été mis en place pour la validation des nouveaux noms de gènes : le Human Gene Nomenclature Committee (HGNC). En parallèle, en 1986, la librairie nationale de médecine des Etats-Unis a lancé le Projet UMLS (Unified Medical Language System) qui avait, entre autres, comme objectif de produire un méta-thésaurus de la terminologie biomédicale (Humphreys et al. 1998). Actuellement, ce méta-thésaurus contient plus de trois millions de termes[30]. Avec le développement de la production scientifique et des bases de données dû à l'évolution des techniques expérimentales, un nouvel outil inspiré de la science de l'information a été utilisé pour la standardisation des terminologies. Il s'agit des ontologies ; contrairement à l'UMLS, les ontologies biomédicales sont des ontologies spécialisées par discipline ; c'est ainsi que le consortium Gene Ontology a été créé pour maintenir Gene Ontology, une ontologie pour la description des fonctions des gènes et leurs produits (Ashburner et al. 2000). D'autres

[29] http://www.genenames.org/about/overview

[30] https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/statistics.html

ontologies ont été développées, pour la description des populations cellulaires (Bard et al. 2005), pour les noms des maladies (Kibbe et al. 2014).

Le portail bioportal est une initiative du Centre National de Bio-Ontologies de l'université de Stanford. Le but est d'offrir un répertoire d'ontologies biomédicales (plus de 540 ontologies y sont répertoriées)[31] et un ensemble d'outils de visualisation, d'alignement et d'annotation pour ces ontologies (Noy et al. 2009). L'initiative Open Biological Ontology (OBO) Foundry (Smith et al. 2007) est aussi une initiative pour répertorier les ontologies biomédicales mais avec des exigences plus strictes pour la standardisation qui ont été adoptées par la communauté. En effet, l'OBO Foundry offre aux ontologistes en sciences biomédicales un modèle de bonnes pratiques et de règles à suivre. Notamment, la définition des relations et des concepts doit contenir un minimum d'indications appartenant à des ontologies déjà publiées dans le portail OBO, le but étant de produire des ontologies interopérables et faciles d'utilisation par des systèmes informatiques. Toutes les ontologies OBO doivent être libres d'accès. Plusieurs outils ont été développés sur la base de ces ontologies notamment pour l'annotation des expériences explorant le génome et le transcriptome.

Cette abondance des données expérimentales, des bases de données biologiques et de la production scientifique, ainsi que l'effort entrepris pour la standardisation et l'interopérabilité, ont permis l'émergence de systèmes experts tirant profit des techniques des sciences de l'information et des technologies de l'exploration des données et des données textuelles pour l'étude des phénomènes biologiques. Dans la section suivante, je vais exposer quelques applications produites pour l'interprétation des données biologiques en utilisant des technologies informationnelles.

3. Approches des sciences de l'information et des sciences de données pour les systèmes biologiques

Dans cette section, je vais présenter deux types d'applications des méthodes de fouille des données (ici on va présenter deux méthodes d'apprentissage automatique) pour le traitement des données de séquençage (MicroARN) et des données post-séquençages (séquences protéiques). La première application est une application médicale ; la deuxième application est une application biologique. Le but est de monter l'intérêt de l'intégration de ces approches pour l'interprétation automatique des résultats d'expériences biologiques. En effet la problématique

[31] http://bioportal.bioontology.org/

sous-jacente du projet de thèse est le développement d'outils d'aide à l'interprétation en adaptant une méthode d'annotation des gènes par contextualisation biologique (population cellulaire, compartiment anatomique et pathologie). La première approche sans pour autant aborder la question directement c'est une approche de contextualisation par pathologie et par compartiment anatomique. L'outil développé est un outil qui permet de reconnaitre seulement 21 types de cancer dans des organes et tissus différents. La seconde approche est une approche qui permet de produire un intéractome de gènes à partir de littérature (une approche qui été utilisée aussi pour le développement de db-STRING que nous utilisons) mais qui ne permet pas justement de situer les gènes dans des voies biologiques intracellulaires.

3.1. Mirna Cancer Analyser [32]

C'est un outil développé par une équipe de l'université de Stanford (Cheerla and Gevaert 2017), pour l'aide à la décision médicale en termes de diagnostic et de pronostic du cancer et même pour l'approche de traitement.

- **Données analysées par Mirna Cancer Analyser** :

  Profil d'expression des MicroARN[33] par tissu/organe et données cliniques

- **Objectifs de développement du Mirna Cancer Analyser** :

Le premier objectif est d'utiliser l'analyse des données d'expression pour préciser le diagnostic et permettre de prédire le pronostic du cancer (type de cancer et quel grade clinique).Le deuxième objectif est de partir de l'analyse des données cliniques pour prédire la réponse du patient en fonction des traitements envisagés.

- **Approche utilisée par Cheerla et Gevaert**

La première étape consiste à déterminer le type du cancer (tissu/organe/origine), le stade et le grade sur la base des données du transcriptome et plus précisément, les données d'expression des micro-ARN. Plusieurs algorithmes de classification ont été testés mais le SVM (Support Vector Machine) est l'algorithme qui a obtenu les meilleures performances. Le SVM est un algorithme de classification supervisée dont le principe est le suivant : on projette dans un hyper espace, les entrées sous forme de vecteurs. L'idée est de trouver des séparateurs linaires entres les vecteurs qu'on veut classer (Dehak et al. 2009).

[32] http://www.mirnanalyze.com/

[33] Les MicroARN : des petits fragments d'ARN composés d'une vingtaine de nucléotides. Ils joueraient un rôle important dans la régulation génétique

L'ensemble d'apprentissage et de validation pour cette étape est composé de trois jeux de données :

- Un jeu de données qui contient des valeurs d'expression de micro-ARN pour plus de 200 cas soufrant de 12 types de cancer.
- Un jeu de données qui contient des valeurs d'expression de micro-ARN pour 8 échantillons de mélanomes
- Un jeu de données qui contient des valeurs d'expression de micro-ARN de 29 échantillons de Lymphome B

Pour l'étape de prédiction du comportement face au traitement, les auteurs se basent sur les données cliniques des patients (âge, genre, groupe ethnique, type de traitement, durée de rémission). Ils commencent par calculer une distance entre les différents patients ayant différents traitements. Cette distance est projetée dans un espace à trois dimensions. Chaque traitement K est associé un point P_K dans cet espace puis la différence entre les distances 3D et les distances réelles est réduite et à la fin du processus les coordonnées 3D de chaque traitement sont consignées. Puis un nouvel algorithme SVM a été développé, combinant les données des micro-ARN, les données cliniques, le type de cancer et les coordonnées 3D du traitement pour prédire la réponse au traitement (rechute ou rémission).

L'ensemble de validation a été constitué à partir des données cliniques de 710 patients de la base de données TCGA.

- **Performances**

Le taux de précision de l'outil est de plus de 97%. Pour la validation et l'entraînement du modèle, les auteurs ont utilisé les données issues de la base de données GEO[34] (Gene Expression Omnibus), une base de données de transcriptomique et la base de données TCGA qui répertorie les MicroARN par tissu normal et par tissu cancéreux. Ces données ont été générées par une plate-forme de séquençage Illumina. L'analyse des données permet de savoir aussi quelle est la réponse potentielle d'un patient à un traitement donné. Avec un taux de précision de 85% ce qui renseigne sur la fiabilité de la technique, l'outil développé améliore considérablement les protocoles de soin. De plus, l'originalité de la technique est qu'elle n'est pas spécifique à un type particulier de cancer mais à vingt-et-un types différents ce qui permet une utilisation répandue. Pour le diagnostic et le pronostics de ces cancers mais aussi pour permettre de proposer des protocoles de soin adapté en fonction du profil d'expression des Micro-ARN.

[34] https://www.ncbi.nlm.nih.gov/geo/

Cette approche ne permet pas par contre de comprendre les mécanismes moléculaires et cellulaires mis en jeux juste d'apporter une aide à la décision pour la prise en charge du médecin.

3.2. Prédiction des interactions protéine-protéine

Cette étude a été présentée par l'université de Vrije à Amesterdam (Hou et al. 2017), l'objectif étant de prédire les interactions protéine-protéine à partir des données génomiques.

- **Problématique liée à la prédiction des sites d'interactions**

La prédiction des séquences et des fonctions protéiques à partir de ces données génomiques et transcriptomiques constitue une étape importante du processus d'interprétation (annotation). Plusieurs outils et base de données ont été développés pour la prédiction des séquences et des fonctions protéiques à partir de l'ADN et de l'ARN[35]. En revanche, le challenge que pose cette abondance de données sur les séquences protéiques est l'identification des protéines qui interagissent ensemble et surtout les sites d'interactions. Les approches expérimentales pour la détermination des interactions protéiques sont coûteuses en termes d'argent et de temps (Jessulat et al. 2011). D'autres méthodes basées sur l'amarrage moléculaire et les études structurales (Stambouli et al. 2014) sont coûteuse en ressources informatiques (calcul et informations).

- **Approche développé par Hou et coll.**

L'approche proposée se base sur un modèle d'arbres décisionnels (voir B.2.2.2) qui permet de combiner cinq paramètres calculés à partir de la comparaison des différentes séquences protéiques avec une courbe ROC dont l'aire est de 0.72.

- **Intérêt de l'apprentissage automatique par rapport aux approches classiques**

L'intérêt de cette méthode est son faible coût (matériel, temps) par rapport aux méthodes expérimentales et aux méthodes basées sur la modélisation 3D. Ce qui nous permet d'imaginer des applications et des processus d'analyse en continu de données génomiques à haut débit. Ici l'approche permet juste de prédire une interaction à partir de la littérature sans autant prédire les voies intracellulaires, c'est-à-dire sans donner un sens biologique à cette information.

[35] predictprotein : https://predictprotein.org/ | transeq : https://www.ebi.ac.uk/Tools/st/emboss_transeq/ | CD search : https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi

4. Approche de fouille des textes et d'annotation textuelle pour la biologie

L'annotation des entités biologiques dans les textes a émergé au cours des années 1990 (Ananiadou et al. 2006). En effet l'augmentation exponentielle de l'offre informationnelle et documentaire scientifique a encouragé l'émergence de nouveaux domaines de recherche scientifique en biologie et en informatique, y compris le *biotext mining* (Ananiadou et al. 2006). Les premiers algorithmes ont répondu aux besoins de l'identification des protéines et des noms de gènes (Denys et al. 1998) et l'annotation des fonctions protéiques (Andrade et Valence, 1998). Dans les années 2000, les approches d'*exact maching* ont été utilisées pour l'annotation biotextuelle, bénéficiant du développement de bases de données stables servant à la construction de dictionnaires, on cite comme exemple la méthode de Krauthammer et Coll. pour l'alignement du texte avec les dictionnaires de noms de protéines et de gènes qui en utilise "BLAST" (Krauthammer et al. 2000). Ces techniques ont l'inconvénient d'être lentes et gourmandes en ressources informatique cependant, elles garantissent de bonnes performances. Plus tard, les techniques d'annotation ont été améliorées en intégrant les algorithmes d'apprentissage automatique pour augmenter les performances d'annotations (Tourouka et al. 2005). L'intégration des méthodes d'apprentissage automatique basées sur les règles d'association et de classification pour la production d'outils hybrides a permis d'aborder de nouveaux problèmes. Cela a conduit à identifier de nouveaux concepts biologiques dans la littérature (Kim et al. 2015), l'extraction d'interactions protéiques (Huang et al. 2004), l'identification de la maladie et des relations génétiques (Bravo et al. 2015 ; Tiffin et al. 2005 ; Kaur et al. 2014), ou des interactions médicamenteuses (Iyer et al. 2014). Ces approches dépendent fortement du corpus d'apprentissage (Jonnalagadda et al. 2012) et sont donc difficiles à généraliser. Plusieurs compétition communautaire pour l'extraction des concepts biologiques à partir des textes ont été organisés pour évaluer ces outils (Krallinger et al. 2008 ; Ananiadou et al. 2015 ; Huang et Lu, 2016). Dans la Table 3 on présente quelques outils de *biotext mining*. Nous avons adopté les applications recommandées dans la littérature. Nous avons choisi un exemple pour chaque approche de developpement (application, site web) et sélectionné des applications qui reconnaissent plusieurs concepts biologiques (gènes, populations cellulaires, processus biologiques) sachant que certaines de ces applications seront utilisées comme références auxquelles l'outil de criblage de la littérature que nous avons développé dans la présente thèse, sera comparé.

D. Conclusion

Pour résumer, nous avons présenté dans ce chapitre, le concept d'information (sa définition et son évolution historique). En biologie, la notion d'informations biologiques fait l'objet d'une controverse qui sera explicité par la suite. Cette controverse est portée par les spécialistes de la théorie de l'information. Une théorie qui aborde le concept d'information d'un point de vue quantitatif et qui fait abstraction du sens. Ce sens qui a été défini par les philosophes de l'information comme dépendant du **contexte**. Or la notion d'information biologique a été adoptée par la communauté biologique depuis la découverte du code génétique. Nous avons aussi présenté les avancés des technologies expérimentales en biologie. Ces avancées sont à l'origine d'un afflux massif d'informations sous forme de textes, de bases de données mais aussi de la production de données expérimentales massives. Ces données produites m'ont poussé à adapter des méthodes d'analyses originales mais aussi à penser des méthodes automatiques pour améliorer leur interprétation. Dans ce qui suit, nous allons présenter une approche inspirée des sciences de l'information et des sciences des données basée sur la notion de **contextualisation** pour l'étude de l'expression des gènes.

Table 3. Table récapitulative des applications les plus importantes d'annotation textuelle pour les textes biologiques.

Nom de l'outil	Auteurs	Type	Concepts identifiés
SciLite	(Venkatesan et al. 2016)	Application Web spécialement pour la base bibliographique PMC européen	Les composées chimiques, les noms de gène, les processus biologiques, les maladies et les organismes
GOTA	(Lena et al. 2015)	Interface Web	Prédiction des processus biologiques, des fonctions moléculaires et des compartiments subcellulaires à partir d'un abstract sur la base de *Gene Ontology*
OBO Annotator	(Groza et al. 2015)	Application JAVA	Annotation des textes sur la base de la *Phenotype Ontology*
MimMiner	(van Driel et al. 2006)	Interface Web	Cette interface permet d'identifier des relations gène-maladie à partir de la base de données OMIM
GoPubMed	(Doms and Schroeder 2005)	Interface Web qui intègre un moteur de recherche pour PubMed et un outil d'annotation	Gene Ontology Les termes MeSH

Chapitre 2 : Annotation Textuelle et Contextualisation (Contribution)

A. Introduction aux données transcriptomiques

Pour conduire ses thématiques de recherche, le laboratoire i3 se base beaucoup sur l'étude des données à haut débit dont transcriptome (c'est à dire l'ensemble des ARN produits à un instant t dans une cellule donnée). Pour cela, nous utilisons les données issues de la technologie des micropuces à ADN (une technologie à haut débit qui permet d'avoir une image instantanée sur l'état de transcription du génome) (Grinberg-Bleyer et al. 2010, Rosenzwajg et al. 2015, Nehar-Belaid et al. 2016). C'est une technologie à haut débit qui nécessite un processus d'analyse statistique conséquent pour l'extraction des observations biologiques et l'annotation et l'interprétation de ces analyses. Dans ce chapitre, je vais commencer par introduire un processus classique d'analyse de données de micropuces. Par la suite, je vais exposer ma contribution pour l'annotation des observations expérimentales et l'aide à l'interprétation des résultats par le développement d'une méthode de contextualisation des gènes et de leurs produits par le criblage automatique de la littérature.

Les techniques d'étude de la transcription à haut débit comme les puces ADN sont très utilisées pour l'étude de l'expression des gènes depuis les années 1990 (Fodor et al. 1991, Lenoir and Giannella 2006). Les premières études ont été menées sur des groupes d'une centaine de gènes (Schena et al. 1995, Anon. 1996). Cette technique a connu un grand essor surtout suite au séquençage du génome humain et à la création d'entreprises spécialisées. Actuellement, les puces produites peuvent cribler un génome entier (Aakula et al. 2015, Riz et al. 2015) ou bien une partie bien spécifique du génome comme les miRNA (Li et al. 2013). Comme pour toute technique à haut débit, les données générées nécessitent un processus d'analyse informatique assez conséquent. Les étapes indispensables peuvent être résumées comme présenté dans ce qui suit (Sarver 2010) :

1. Contrôle qualité

Avant d'entamer les étapes d'analyses statistiques et de normalisation, il faut, comme pour toute expérience scientifique, séparer le bruit de fond du « vrai signal ». On pourrait résumer cette étape sous forme de questionnaire :

- Est-ce qu'il y a une distribution uniforme des signaux des sondes, ou bien observe-t-on une variation du signal pour certaines expériences ?

S'il y a une variation significative entre les différentes puces, on peut éliminer les échantillons avec un très faible signal. Une méthode très utilisée pour visualiser la distribution du signal est la méthode des *boxplots*.

- Est-ce que l'intensité globale du signal est suffisante pour déterminer des différences significatives de transcription ?

  Ceci peut être estimé par comparaison avec les résultats d'autres puces, qui ont été précédemment analysées avec succès et se sont avérées informatives.

- Les réplicats techniques et les réplicats biologiques sont-ils plus corrélés entre eux que les échantillons des différentes expériences ?

  Si des échantillons provenant de répliques biologiques ne montrent pas une corrélation plus importante que les échantillons provenant de différents groupes, un problème systématique existe avec l'expérience et les analyses statistiques ne devraient pas être effectuées. Une mesure de corrélation de Pearson (Soper et al. 1917; Jaffe et al. 2017) peut être effectuée à cette étape.

- Est-ce que les conditions expérimentales peuvent être séparées par des méthodes automatiques ?

2. Filtrage des données

Cette étape a pour but d'éliminer le bruit de fond. Plusieurs approches sont utilisées pour le filtrage des données. Par exemple Tseng et coll. dans une étude menée sur une puce ARN dichrome de 125 gènes d'*Escherichia coli* considèrent les paramètres suivants pour filtrer (Tseng et al. 2001) :

- Ils calculent le rapport $m = \frac{Cy5}{Cy3}$[36] pour chaque sonde.
- Chaque gène ayant plusieurs sondes ils calculent le rapport $CV = \frac{écart\ type}{moyenne}$ de toutes les sondes pour chaque gène car le rapport CV est corrélé avec la qualité de la sonde.
- Pour chaque gène cible, ils considèrent un sous groupe de 50 gènes dont les moyennes d'intensité de fluorescence sont proches de celle du gène cible.

[36] Ici on note Cy3 et Cy5 en référence à la Cyanine utilisée pour le marquage des sondes. En effet on utilise le Cy3 pour le marquage rouge et Cy5 pour le marquage en vert. Ici cette notation indique le taux de fluorescence.

- Si le CV du gène cible n'est pas dans les 10% les plus haut du sous-groupe, ses valeurs de fluorescence sont considérées comme bruit de fond et il est ainsi éliminé.

D'autres approches statistiques ont été utilisées pour le filtrage des données comme l'algorithme MAS5 (McClintick et Edenberg 2006) qui se base sur un « test statistique non-paramétrique », le test de de rang de Wilcoxon, développé pour les puces *affymetrix*®. A noter que l'étape de filtrage et de normalisation peuvent être conduite simultanément.

3. Etape de normalisation

Une expérience d'étude du transcriptome par les micro-puces peut être constituée de plusieurs puces. Pour pouvoir comparer le signal issu de plusieurs puces, il faut rendre des valeurs comparables. Des variations peuvent être dues aux conditions de l'expérimentation (température, quantité du fluorochrome qui peut varier…). Ces techniques étant très sensibles, un biais peut être introduit assez facilement. On passe alors par une étape de normalisation pour rendre les signaux issus de toutes les puces comparables. Dans un article publié dans la revue *Briefings in Bioinformatics*, la normalisation est présentée comme un processus destiné à diminuer les variations non biologiques (Kreil and Russell 2005). Plusieurs approches sont proposées pour répondre à cet objectif. Nous citons ici deux types d'approches à titre d'exemple :

- Le rééchelonnement de l'intensité du signal par la médiane : Cette méthode suppose que tous les échantillons partagent la même médiane. C'est une méthode simple mais qui « *donne une robustesse aux valeurs aberrantes* » (Kreil and Russell 2005). Une des variantes de cette méthode est le rééchelonnement par la médiane de l'intensité du signal des gènes de ménage[37] car on suppose que leur expression est constante (Lemma et al. 2016). Bien que certaines études démontrent que certains de ces gènes sont différentiellement exprimés (Kalendar et al. 2017). Une autre approche se base sur l'utilisation de l'ARN pointe[38] pour l'étape de normalisation à la place ou en complément de l'utilisation des gènes de ménage.

[37] Gènes qui sont toujours exprimés. Ils sont aussi nommés : gènes d'entretien ou gènes domestiques. Ils sont réputés s'exprimer de la même manière dans toutes les cellules d'un organisme. Leurs produits d'expression sont indispensables à la vie de la cellule.

[38] En anglais c'est le spike RNA, c'est de l'ARN dont la quantité est connue qui est ajouté au début de l'expérience avant l'étape de l'extraction.

- La normalisation par régression (exemple : la méthode LOWESS) : LOWESS est l'abréviation de *locally weighted linear regression*. Cette méthode a été optimisée pour les expériences de transcriptomique utilisant deux fluorochromes : vert dont l'intensité est notée (Gi) et rouge dont l'intensité est notée (Ri). C'est une relation de régression entre le rapport $\boldsymbol{log_2(\frac{R_i}{G_i})}$ et $\boldsymbol{log_{10}(Ri * Gi)}$ (Quackenbush 2002). La méthode a été développée dans le but de détecter et corriger les écarts dans un diagramme R-I (*Figure 24*).

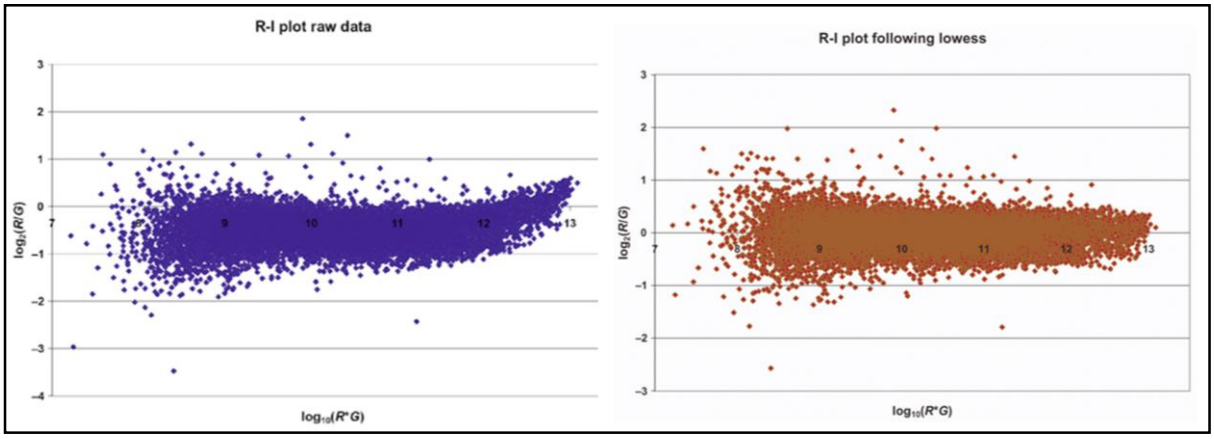


Figure 24. Diagramme R-I représentant la relation : $log_2\left(\frac{R_i}{G_i}\right) = f(log_{10}(R_i * G_i))$.
Les données présentées ici sont des données issues d'une micropuce de souris contenant 27 642 éléments. A gauche, les points bleus représentent les données non normalisées. A droite, les points marron présentent les données normalisées par LOWESS. Source : (Quackenbush 2002)

4. Analyse des données

Le but de cette étape d'analyse est de déterminer les gènes exprimés différentiellement en fonction des conditions expérimentales définies par l'expérimentateur. Après s'être assuré de la qualité de l'expérience, des tests statistiques basiques peuvent être effectués tels que le t-test ou l'ANOVA pour l'analyse de la variance. Pour s'assurer que les observations sont biologiquement significatives, on se base sur les p-values déterminées par ces tests. D'autres tests statistiques et comparatifs peuvent être effectués pour quantifier la sur-expression et la sous-expression de gènes d'une expérience à une autre. Dans ce sens, plusieurs outils ont été développés (Saeed et al. 2003, p. 4, Grace and Nacheva 2012, Calabrese and Mario 2015). Une étape de regroupement est indispensable pour identifier les gènes avec le même profil d'expression.

Une étape de regroupement est indispensable pour identifier les gènes avec des profils d'expression proches dans différentes conditions expérimentales. Des approches simples peuvent être utilisées pour ce regroupement dont à titre d'exemple le diagramme de Venn

(Aumsuwan et al. 2016). Par ailleurs, des algorithmes de regroupement sont aussi rapportés dans la littérature telle que le bi-clustering (McLachlan et al. 2017), le clustering hiérarchique (Wang et al. 2011) et l'algorithme des k-means (Calvano et al. 2005). Plus récemment, des méthodes plus complexes basées sur la fouille des données et l'apprentissage automatique et statistique sont décrites comme les réseaux bayésiens (Wu and Zhang 2016) ou les algorithmes génétiques (Paul et al. 2016).

5. Une étape d'annotation

Après ce processus d'analyse qualitative et quantitative, il s'agit d'interpréter les résultats obtenus. On passe ici de l'étape d'analyse statistique à l'analyse de l'information. C'est le passage de l'observation expérimentale (la quantification) à l'interprétation biologique. Ce passage se fait par confrontation des données expérimentales aux bases de connaissances biologiques. Pour ce type d'expériences, il s'agit des bases spécialisées en biologie fonctionnelle des gènes : les ontologies, en particulier *Gene Ontology* (Ashburner et al. 2000), ou bien les bases de connaissances sur les réseaux biologiques telles que KEGG pathways (Kanehisa and Goto 2000). La technologie utilisée ici étant une technologie à haut débit et le nombre et la taille des bases de données étant importante, plusieurs outils informatiques ont été développés pour l'annotation et l'aide à l'interprétation des données biologiques. On peut citer DAVID (Dennis Jr et al. 2003), un site web de référence qu'on peut aussi interroger via des API[39]. DAVID est un outil qui permet d'interroger plusieurs bases de connaissances dont KEGG et Gene Ontology et permet d'annoter les gènes issus des expériences et de voir ainsi les processus et les voies biologiques impliquées. Ingenuity Pathway Analysis (IPA®) (Krämer et al. 2014) est une autre solution maintenue par une entreprise et qui a été développée pour l'annotation des données à haut débit. Contrairement à DAVID, IPA intègre sa propre base de connaissances issue, outre des bases de données publiques, des données issues du criblage semi-automatique de la littérature et des algorithmes prédictifs. La MSigDB® (Molecular Signature Database) (Subramanian et al. 2005) est une base de connaissances contenant des groupes de gènes (signatures) pré-annotés par un criblage manuel de la littérature des bases de données et de *Gene Ontology*.

Dans ce contexte le laboratoire a produit une stratégie innovante pour l'analyse statistique des puces transcriptomiques (Pham 2013, Pham et al. 2014, Dérian et al. 2016) et qui été présenté

[39] Ce sont des fonctions par lesquelles un logiciel offre des services à d'autres logiciels

dans une thèse présidente (Pham 2013), c'est un teste statistique d'analyse en composantes indépendantes suivie d'un teste d'identification de signatures GSEA.

B. Objectif de notre approche pour la contextualisation

Comme mentionné plus haut, le laboratoire i3 utilise des données issues de micropuces et une des problématiques concerne l'étude du vieillissement du système immunitaire, un système biologique complexe avec une organisation multi-échelle. En effet, ce système est caractérisé par l'intervention d'une multitude de voies biologiques impliquant plusieurs molécules (les gènes et leurs produits…), et une variété de populations cellulaires intervenant dans plusieurs organes et compartiments de l'organisme. A titre d'exemple, une étude multiparamétrique a été menée en utilisant l'approche puce ADN à haut débit pour l'étude du transcriptome de lignées souris à plusieurs âges (Pham 2013). Une approche pour l'identification de signatures de gènes avec une expression différentielle a été développée et utilisée pour l'analyse de la puce (Pham et al. 2014). L'utilisation des approches classiques (Dennis Jr et al. 2003, Krämer et al. 2014) pour l'annotation de ces signatures a montré ses limites. Une partie des limites peut être expliquée par la redondance des sources d'annotation. En effet, les trois outils cités dans le paragraphe précédent utilisent KEGG et *Gene Ontology* dans leur base d'annotation. Or, dans *Gene Ontology*, seulement 63 concepts décrivent des processus du système immunitaire et KEGG intègre 20 réseaux immunitaires seulement. La base de données MSigDB et l'outil IPA contiennent des signatures issues de l'analyse de la littérature mais qui est criblé de facons manuelle ou semi-automatique donc ce sont des approches à bas débit et qui tiennes pas ou peu compte du contexte d'expression. En effet, la littérature constitue une source d'annotation très riche et très peu exploitée. Plusieurs auteurs suggèrent de lier les données de micropuces à la littérature (Masys 2001, Krallinger et al. 2005), mais les approches les plus développées ont été des approches non supervisées qui se sont intéressées à la déduction des fonctions biologiques (Oliveros et al. 2000, Blaschke et al. 2001). Nous avons donc décidé de développer une approche supervisée basée sur l'étude de la littérature pour la production de réseaux de gènes contextualisés. Les notions de contexte et de contextualisation ont été largement abordées en linguistique et en didactique. Ainsi Blanchet définit le contexte comme « *Ce vers quoi ou sur quoi ne convergent pas la focale / ce qui n'est pas au centre de la focalisation / ce depuis quoi on règle les focales (y compris le contexte du chercheur et de la recherche) mais qu'on fait néanmoins entrer dans le champ* » et la contextualisation «*Attribuer des significations à des phénomènes sur lesquels on focalise l'observation, phénomènes qu'on inscrit dans le*

continuum des pratiques sociales en mobilisant d'autres phénomènes qu'on choisit de faire entrer dans le champ au titre de contexte (de paramètres contextuels efficients) mais qui ne sont pas au centre de la focale » (Blanchet 2012). Ainsi pour déchiffrer le sens d'une information il est indispensable de la mettre dans son contexte. Dans le cas de l'information génétique, le sens est apprécié par sa fonction au sein d'une voie biologique (la notion de voies biologiques sera abordée dans le chapitre suivant). Cette fonction ne s'exprime que sous certaines conditions : on peut citer l'environnement intra-cellulaire, épigénétique, environnement extracellulaire, variables individuelles... Depuis l'expérience de Jacob et Monod et l'émergence de la biologie développementale nous savons que des facteurs régulent l'expression des gènes. Ainsi, l'expression des gènes est conditionnée par un contexte d'expression «favorable ». Dans le présent travail, nous nous sommes focalisés sur trois paramètres qu'on a jugés observables avec les méthodes actuelles, pour définir le contexte d'expression : le contexte cellulaire, anatomique et pathologique.

La première étape consiste à produire des groupes de gènes co-cités dans le même contexte d'étude à partir de la littérature vue que durant revu de la littérature nous n'avons pas trouvé de d'outils capables de procéder à cette étape de contextualisation à grande échelle. Cette première étape de contextualisation suivie de l'étape de modélisation des voies biologique (présentée dans le chapitre 3) permettra de produire une base de connaissances sur les voies biologique mis en jeux par contexte que nous pourrons utiliser pour annoter les données de transcriptomique sur le vieillissement du système immunitaire. Nous définissons un contexte d'étude par une population cellulaire, une localisation anatomique (tissus, organe, compartiment…) et/ou une pathologie. En effet, la littérature démontre que l'expression des gènes varie en fonction de la population cellulaire (Zhang et al. 1997, Palmer et al. 2006, Li et al. 2016), du tissu (Nishimura et al. 2004, Wong et al. 2016) et de l'état pathologique (Yang et al. 2014, Calon et al. 2015). Par conséquent, les processus biologiques mis en jeux d'un contexte à un autre devraient être différents. Nous avons donc commencé par développer un outil, baptisé OntoContext, qui permet d'annoter les textes en fonction de ces trois paramètres et d'identifier les noms des gènes dans ces contextes. Notre but était de fouiller un grand nombre d'abstracts de la base de données MEDLINE (Smith et al. 1992) issus d'une recherche supervisée. Pour interroger cette base de données, nous avons lancé plusieurs requêtes dans MEDLINE en utilisant les options avancées de PubMED. Cette requête se fait moyennant les « *MeSH terms* ».

- MEDLINE (Smith et al. 1992) est une base de données bibliographique spécialisée dans les sciences de la vie et l'information biomédicale. Elle contient des citations et des résumés

d'articles couvrant plusieurs champs disciplinaires allant de la médecine, l'infirmerie, la médecine dentaire, les sciences vétérinaires et la santé en général à la biologie et la biochimie… Elle est maintenue par la Bibliothèque Nationale de Médecine des Etats-Unis (NLM), disponible gratuitement sur Internet. On peut l'interroger assez facilement via le moteur de recherche PubMED. En 2015, elle contenait plus de 23 millions de références (citation + résumé) rédigées à partir de l'année 1950.

- PubMED (Lu 2011) est un moteur de recherche de données bibliographiques développé par le centre national américain pour les informations en biotechnologie (NCBI), pour interroger la base de données MEDLINE. Il est gratuit et libre d'accès.

- MeSH terms (Sewell 1964) est un vocabulaire biomédical standardisé hiérarchisé en langue anglaise, utilisé comme système d'indexation par le NLM et le NCBI pour les références bibliographiques de la base de données MEDLINE.

Nous avons décidé de prendre l'exemple du vieillissement du système immunitaire pour tester notre approche vu les défis qu'il présente. En effet l'information disponible sur les processus impliqués dans le vieillissement dans les sources d'annotations classiques est rare dans *Gene Ontology*. Le terme « *Aging* » a été introduit depuis 2001[40]. Actuellement il y 320 produits de gènes humains qui sont annotés par ce terme, dont seulement 92 ont été produits expérimentalement. Dans la base de données KEGG on compte 3 voies biologiques en relation avec le vieillissement. Le système immunitaire quant à lui est un système complexe : multi échelle, dynamique et fluide. Dans *Gene Ontology* plus de 8000 produit de gènes humains sont annotés par le terme *Immune system process* et 20 voies biologiques sont répertoriées dans KEGG. Néanmoins aucune de ces deux sources ne répertorie soit des termes ou bien des voies biologiques sur le vieillissement du système immunitaire, de plus l'organisation multi échelle du système immunitaire nous pousse à bien différencier les gènes et leurs produits en fonction des populations cellulaires et les compartiments anatomiques. Son rôle dans la préservation de l'intégrité de l'organisme nous pousse aussi à différencier les pathologies associées. La base de données Gene Card (Stelzer et al. 2016) quant à elle répertorie pour chaque gène un certain nombre d'information (Réseaux biologique, fonction biologiques, pathologies liées…) mais là aussi on se trouve face à la même problématique de redondance des sources puisque c'est les

[40] http://www.ebi.ac.uk/QuickGO/GTerm?id=GO:0007568#term=history

mêmes bases de données qui sont utilisées (KEGG, Gene Ontology…) et donc une limites quant à l'étude des gènes du système immunitaire vue le manque d'annotations pour ce système.

C. Développement d'OntoContext, un package Python pour l'annotation des textes

L'objectif de l'approche est présenté dans la Figure 25. Les étapes de développement d'OntoContext font l'objet d'une publication soumise et présentée en annexe. Dans ce paragraphe est présenté un résumé de cet article.

OntoContext est un package Python[41]. Nous avons choisi de développer un package, vue la masse de données que nous voulons traiter (à titre d'exemple dans notre expérience sur le vieillissement, nous traitons plus de 120 000 résumés d'articles). Le choix du langage de programmation Python est justifié par le fait de sa grande utilisation dans la biologie et la bioinformatique (Bassi 2007).

Le package que nous avons développé comprend deux modules : un modules *Annot*, dont la conception est présentée dans la Figure 26. Comme son nom l'indique, ce module a été développé dans le but d'annoter des documents textes avec trois ontologies :

- L'ontologie *Cell Ontology* pour les populations cellulaires (Bard et al. 2005)
- L'ontologie *UBERON Ontology* pour les localisations anatomiques (Mungall et al. 2012)
- L'ontologie *Human Disease Ontology* pour les concepts pathologiques (Kibbe et al. 2014)

Pour l'annotation textuelle nous avons commencé par produire trois tables pour chaque ontologie :

- Une table synonymes qui permet de répertorier pour chaque concept les synonymes et les pluriels de chaque concept.
- Une table étiquettes qui permet de répertorier pour chaque terme (concept ou synonyme) sont étiquette morphosyntaxique.
- Une table des chemins qui répertorie pour chaque concept l'ensemble de cet ancêtre qui va nous permettre par la suite de regrouper les articles ou les concepts et leurs termes fils ont été cité.

[41] https://www.Python.org/

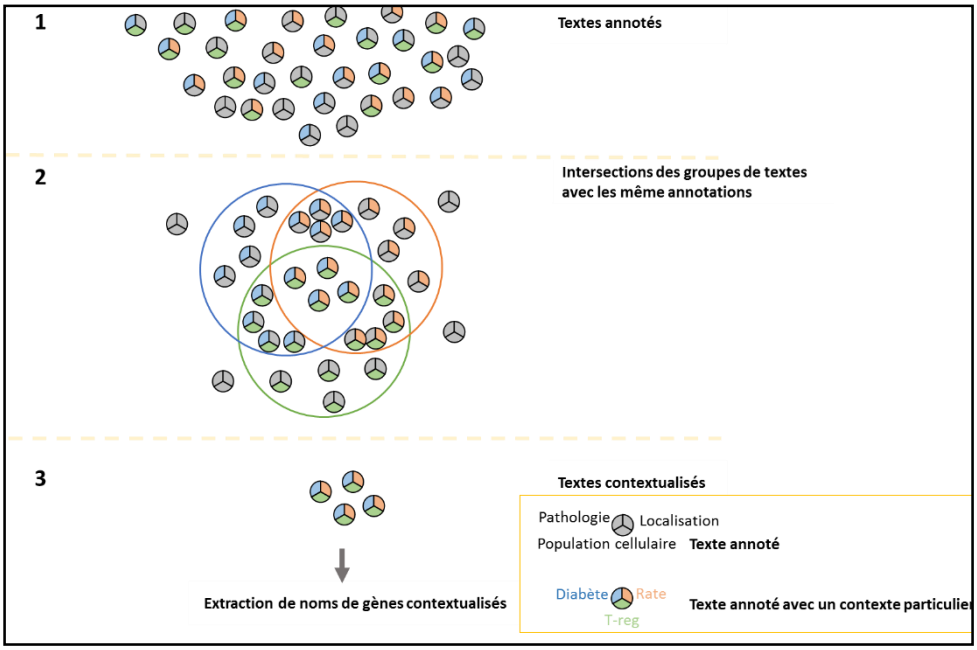


Figure 25.Objectif de l'approche. 1) On commence par annoter un corpus de textes (ou bien des résumés issus de PubMED. 2) On définit un contexte bien particulier (ici Rate, Diabète, et T-reg). 3) Les textes annotés avec les trois termes de notre contexte sont sélectionnés pour l'identification des noms de gènes et de leurs produits.

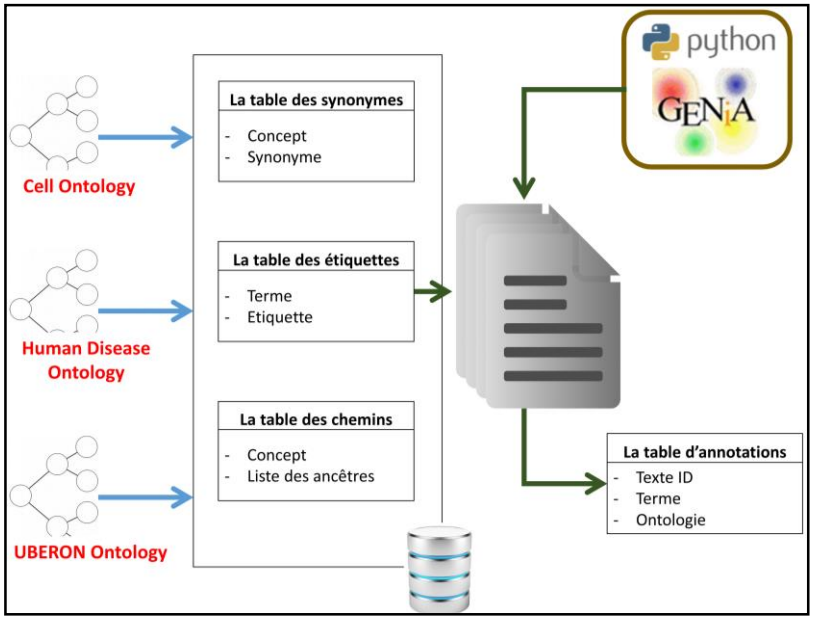


Figure 26. Conception du module *Annot* du package OntoContext. Les flèches bleues constituent un prétraitement que nous avons effectué à partir des trois ontologies sélectionnées pour la production de notre base d'annotation. La base d'annotation contient trois tables pour chaque ontologie (synonymes, étiquettes, chemins). Les étiquettes sont des étiquettes morphosyntaxiques (POS-Tagging) générées grâce au package NLTK[42] de Python. Les flèches vertes correspondent au déroulement du module Annot. Le module prend en entrée une liste de documents textes. Il procède à la compartimentation du texte, son étiquetage morphosyntaxique pour l'identification des concepts cellulaires anatomiques et pathologiques. Le package GENIA TAGGER est utilisé pour l'identification des noms de gènes

[42] http://www.nltk.org/

et de leurs produits. En sortie la fonction *Annot* génère une table avec les concepts identifiés pour chaque article (annotation).

Au sein du même package nous avons développé un second module que nous avons nommé *crisscross*. Ce module propose une interface graphique qui donne la possibilité à l'utilisateur de choisir un contexte ou des contextes d'étude bien particuliers pour assurer un confort d'utilisation.

1. Etapes de développement

Durant cette partie je vais présenter brièvement les différentes étapes et le principe de développement d'OntoContext. Cette partie fait l'objet d'une publication en cours de soumission (Voir Article soumis).

1.1. Matériel utilisé

Nous avons décidé de développer un package Python pour faciliter l'utilisation et l'intégration de l'outil dans d'autres codes, notamment pour permettre de paralléliser les calculs par la suite vu le volume d'information qu'on veut traiter.

Le package produit a été développé en utilisant la version 2.7 de python. Il a nécessité l'utilisation des packages suivant :

- NLTK [43]: pour l'étiquetage morphosyntaxique
- GENIATAGGER 1.0 Package[44] (Tsuruoka et al. 2005) : pour l'identification des noms des gènes et leurs produits (ARN, Protéine)
- Sqlite3[45] : pour la gestion des bases de données

La package est publié en ligne : https://github.com/walidbedhiafi/OntoContext1, et dans le répertoire pypi pour une plus grande facilité d'utilisation et d'installation : https://pypi.python.org/pypi/OntoContext/0.11.

1.2. Le module d'annotation

Le but de ce module est d'annoter des textes à haut débit en fonction des trois ontologies citée en haut. A partir de ces ontologies nous avons extrait les concepts leurs synonymes (termes), nous avons développé une méthode pour générer automatiquement les

[43] http://www.nltk.org

[44] http://www.nactem.ac.uk/GENIA/tagger/

[45] https://docs.python.org/2/library/sqlite3.html

pluriels des concepts. Nous avons utilisé le package NLTK pour la production des étiquettes (patrons) morphosyntaxiques des termes. Nous avons ainsi produit une base de données avec des tables contenants les concepts et leurs synonymes, les concepts et leurs fils et les termes avec leurs étiquettes en utilisant le package NLTK. La base de données est livrée avec le package.

L'étiquetage morpho-syntaxique (*part-of-speech tagging*) est un processus informatique qui consiste à associer les fonctions gramaticales aux mots dans un texte.

Le principe d'annotation est basé sur l'exact matching, et l'utilisation l'étiquetage morphosyntaxique est présentée dans la Figure 27. On utilise le package NLTK pour morceler et étiqueter les termes des dictionnaires (Figure 27 A) et les phrases des textes (Figure 27 B). Après la compartimentation et l'étiquetage des phrases (Etape1), OntoContext cherche des patrons morphosyntaxiques en commun entre la phrase et la base de données (Etape2 Figure 27). A partir de ces étiquettes OntoContext identifie les termes qui correspondent aux termes de la base de données (Etape 3 Figure 27 et Etape 4 Figure 27). Ce processus est itéré pour chaque terme dans la base de données (Etape 5 Figure 27). Pour récupérer les noms de gènes on utilise le package *geniatagger* 0.1 (Tsuruoka et al. 2005) qui permet d'effectuer un étiquetage morphosyntaxique des textes biomédicaux-avec la particularité d'attribuer des étiquette spécifiques pour les concepts biologique (Protein : pour les protéines, RNA : pour les ARN, DNA : pour les ADN). Les annotations sont stockées dans une nouvelle table. Cette étape est effectuée par la fonction "*annot*" du module "*annotation*".

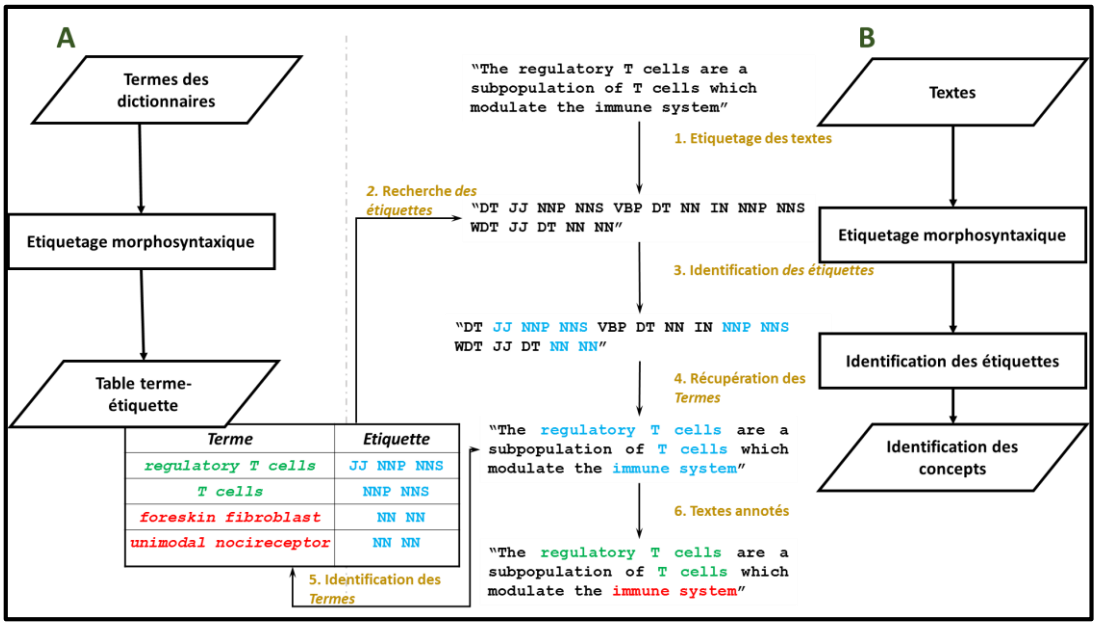


Figure 27. Principe de l'algorithme d'annotation. (A) les concepts et leurs synonymes sont extraits des ontologies et les pluriels sont générées. Ces termes sont étiquetés avec le package NLTK. Après cette étape nous générons la table terme-étiquette. Ici nous montrons un exemple de cette table avec une colonne pour les termes et une colonne pour les étiquettes. (B) 1/Les textes d'intérêts sont parcourus et étiquetés en utilisant le package NLTK. 2-3/Les étiquettes qui correspondent à la table terme-étiquette sont identifiées (en bleu). 4/ Les expressions correspondantes sont récupérées (en bleu). 5/ Ces expressions récupérées sont utilisées pour interroger la base de données des termes. 6/Les expressions qui correspondent à la table Terme-Etiquette sont retenus (En vert) et en rouge les termes qui ne sont pas reconnus. "DT", déterminant ; "JJ", adjectif ; "NNP", Nom propre singulier ; "NNS", Nom au pluriel ; "VBP", Verbe au présent ; "NN", Nom commun au singulier ; "WDT", Wh-déterminant.

1.3. Le module d'entrecroisement

Le module d'entrecroisement est un module graphique offert à l'utilisateur pour un confort d'utilisation (Figure 28). Ce module offre une interface développée avec le package Tkinter[46] de python. L'interface développée affiche tous les termes reconnus ontologie par ontologie en faisant l'extension aux ancêtres. L'utilisateur ici peut choisir une manière de trier les termes identifié soit par fréquence de mention dans le corpus soit par ordre alphabétique. Ainsi il pourra choisir pour chaque ontologie les termes constitutifs de ces contextes d'étude. En fin de processus l'utilisateur aura des listes de noms de gènes identifiés regroupées par contexte d'étude.

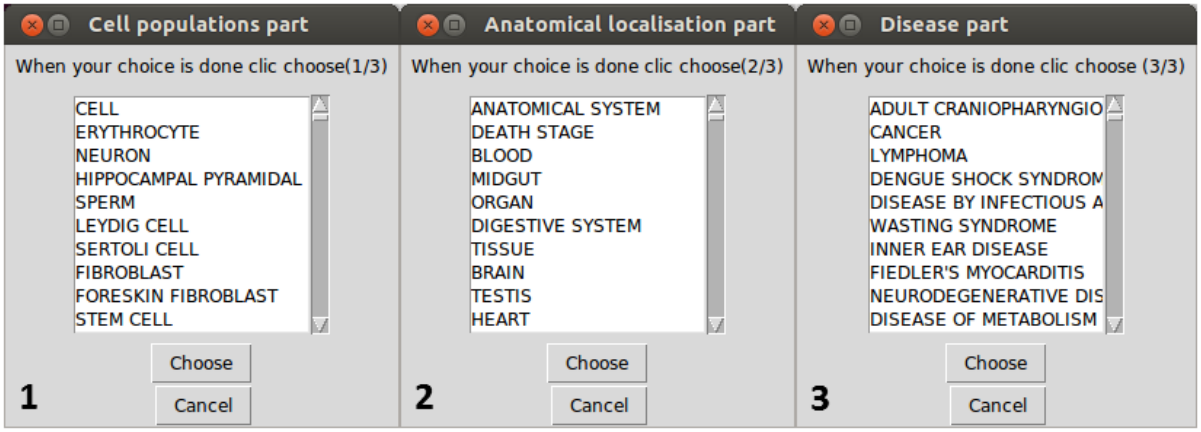


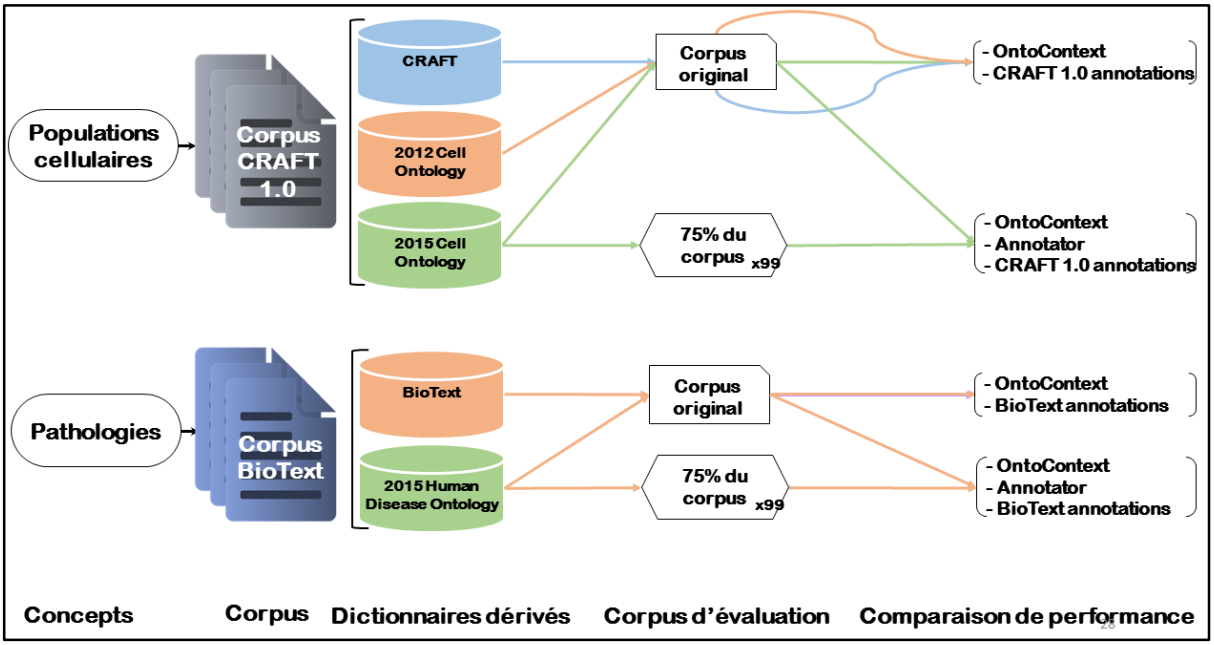
Figure 28. Aperçu de l'interface graphique du module crisscros pour l'entrecroisement. 1. L'interface qui permet de choisir les concepts cellulaires. 2. L'interface qui permet de choisir les concepts anatomiques. 3. L'interface qui permet de choisir les concepts pathologiques.

2. Approche de validation

Pour la validation de l'outil, nous nous sommes focalisés sur la validation de l'outil *Annot* et nous avons confronté le résultat de l'annotation par OntoContext à des corpus annotés

[46] https://wiki.python.org/moin/TkInter

manuellement pour deux types de concepts. Nous avons choisi deux corpus de textes publiés et nous avons confronté les performances d'annotation OntoContext aux textes annotés manuellement utilisés comme référence (Tableau supplémentaire S1 dans la partie article



soumi). Le Corpus CRAFT 1.0 annoté par des experts pour les concepts de population cellulaire est un ensemble de 67 textes complets (Bada et al. 2012). Les annotations manuelles sont basées sur la *Cell Ontology* dans sa version de 2012. Le corpus BioText annoté par des experts pour les concepts pathologiques comprend 141 résumés (Rosario et Hearst, 2004). Nous n'avons pas validé OntoContext pour des concepts anatomiques en l'absence d'un corpus de référence. Les tests de validations sont présenté dans la Figure 29.

Figure 29. Processus de validation du module d'annotation d'OntoContext. Les concepts cellulaires et pathologiques ont été validés chacun par un corpus correspondant annoté manuellement. Des dictionnaires dérivés des annotations manuelles ont été produits et des dictionnaires issues des ontologies et ont été utilisés pour la validation. 35 textes du corpus CRAFT 1.0 et 100 du corpus BioText ont été utilisés pour une comparaison avec l'outil NCBO Anotator. Par la suite un échantillon aléatoire de 75% des textes des corpus a été utilisé pour une deuxième étape de validation et on a comparé à chaque fois l'annotation OntoContext et l'annotation NCBO annotator, on considère l'annotation manuelle comme le vrai positif, et on a répété cette opération 99 fois tout en modifiant les textes à chaque fois.

Pour les concepts cellulaires, nous avons trouvé une précision[47] moyenne de 78%, un taux de rappel[48] moyen de 60% en utilisant l'ontologie *Cell Type* dans sa version de 2015 et en

[47] $précision = \frac{Vraies\ Positifs}{Vraies\ positifs + Faux\ Positifs}$

[48] $rappel = \frac{Vraies\ Positifs}{Vraies\ positifs + Faux\ Négatifs}$

utilisant le corpus CRAFT 1.0 (Bada et al. 2012). Pour les concepts pathologiques nous avons utilisé le corpus BioText (Rosario and Hearst 2004) et l'ontologie *Human Disease Ontology* dans sa version de 2015 et nous avons calculé un taux de précision moyen de 70% et un taux de rappel de 29%. Nous avons soumis l'outil NCBO Annotator, publié par l'université de Stanford (Jonquet et al. 2009), aux mêmes tests d'annotation et avons trouvé des taux de précision et de rappel de 82% et 23% pour les populations cellulaires et de 59% et 21% pour les concepts pathologiques montrant ainsi que notre approche d'annotation est plus performante pour l'identification des concepts. Le détail de la méthode d'annotation et des tests de validation est présenté dans l'article soumis présenté en fin de ce manuscrit.

D. Utilisation d'OntoContext pour le vieillissement du système immunitaire

Comme nous l'avons mentionné précédemment, le but de notre étude est l'identification de processus biologiques impliqués dans le vieillissement du système immunitaire. Pour cela, nous avons décidé d'annoter un corpus de résumés de PubMed qui traite du vieillissement afin d'identifier des listes de gènes et de leurs produits contextualisées. Ceci nous permettra de produire une base de connaissances plus précisément une base de signatures de gènes contextualisées (contrairement au vocabulaire utilisé en transcriptomique, nous utilisons le terme signature ici pour désigner une liste de gènes cités dans le même contexte). Le processus est résumé dans la Figure 30.

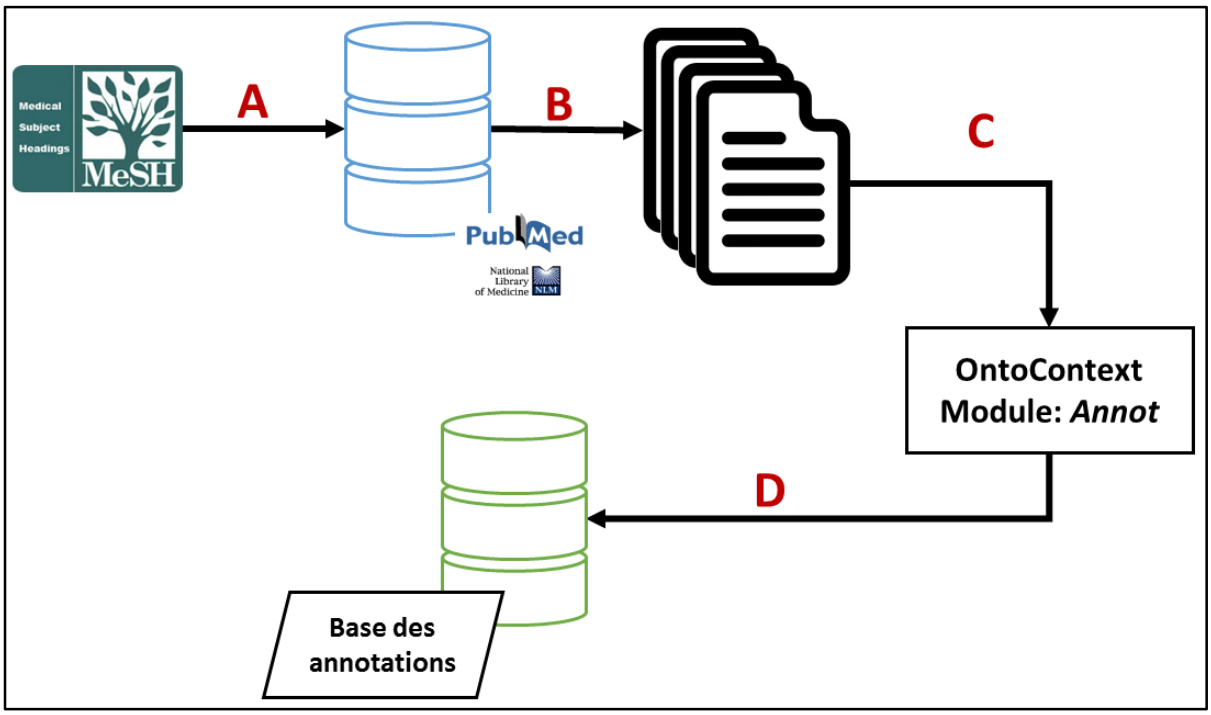


Figure 30. Utilisation d'OntoContext pour le criblage d'un corpus de textes. A. La requête a été construite en combinant une liste de termes MeSH. B. Cette requête a été lancée dans la base de données MEDLINE via l'outil PubMed. C. La liste des résumés d'articles générée est criblée en utilisant le package OntoContext de python. D. A la fin du processus une base de données des annotations cellulaires anatomiques et pathologiques est générée.

1. Choix de la requête PubMed

Pour le choix de notre corpus cible portant sur l'étude du vieillissement du système immunitaire, nous avons fait appel à l'expertise des deux laboratoires. Nous avons convenu de construire notre requête autour des mots clés suivants : *immune system* + *aging* et synonymes (*old, elderly, age, longevity*) + *human* ou *mice* ou *mus musculus* + *gene* ou *protein* ou *molecule*.

Pour la construction de cette requête, nous nous sommes basés sur le vocabulaire MeSH. Un des avantages de ce vocabulaire est son organisation hiérarchique. Nous avons procédé à une étude préalable pour choisir les termes précis :

- *Aging* : (http://www.ncbi.nlm.nih.gov/mesh/68000375), hiérarchiquement ce terme est un terme parent de *Longevity* sous la catégorie « Phenomena and Process category », donc sémantiquement son utilisation implique implicitement le terme *Longevity*.
- *Old* et *Elderly* : ces deux termes n'existent pas dans le vocabulaire MeSH ; en revanche, les termes *Frail Eldery* (http://www.ncbi.nlm.nih.gov/mesh/68016330) et *Aged* sont hiérarchiquement des termes fils de *Age Group* caractérisé sous « Persons Category ».

- *Molecule* : ce terme n'existe pas non plus dans le vocabulaire MeSH ; peut-être est-ce un terme trop générique pour être utilisé par le system MeSH. Nous avons donc décidé de le remplacer par le terme *Antigens* (http://www.ncbi.nlm.nih.gov/mesh/68000941) vu l'intérêt de l'étude pour le système immunitaire.
- *Mus musculus* : dans le système MeSH, il s'agit d'un synonyme du terme *Mice* (http://www.ncbi.nlm.nih.gov/mesh/68051379) ; le terme *Mice*, seul, sera donc utilisé pour cette requête.

Pour résumer, notre requête pour PubMed est la suivante :

((aging OR age groups) AND (proteins OR genes OR antigens) AND (human OR mice) AND immune system)

Tous les termes utilisés ici sont des MeSH terms. Nous avons choisi de lancer la recherche pour l'humain et la souris car on considère que dans la majorité des études publiées, la souris est utilisée comme modèle d'étude pour des applications humaines.

Nous avons lancé cette requête le 29 avril 2016 et avons obtenu 134 041 résumés répertoriés par PubMed depuis 1950. Nous avons alors décidé de filtrer les articles à partir des années 1980. Ainsi, nous avons sélectionné un corpus de 123 393 résumés pour notre étude que nous appellerons corpus du vieillissement du système immunitaire (CVSI) par la suite. Nous avons pris uiquement les résumées ici car ils sont disponibles gratuitement.

2. Analyse des résultats de l'annotation du corpus CVSI

L'utilisation d'OntoContext pour l'annotation du corpus du CVSI nous a permis d'extraire :

- 2313 concepts cellulaires sur la base de la Cell Type Ontology 2015
- 476 concepts anatomiques sur la base de l'UBERON Ontology 2105
- 3695 concepts pathologiques sur la base de la Human Disease Ontology 2015

Des résultats plus détaillé sont présenté en annexe (de 1 à 6).

- **Etude des concepts cellulaires**

  Seulement 2313 concepts ont été utilisés sur 8523 termes de notre dictionnaire de concepts cellulaires. La majeure partie des concepts sont des populations de cellules

immunitaires. Le concept le plus cité est *T CELL* (Figure 31). Il a été cité dans plus de 35 000 résumés articles.

- **Les concepts anatomiques**

  L'ontologie UBERON comporte des concepts qui se réfèrent à des tissus, des organes, des compartiments et des systèmes biologiques. Il y a seulement 473 concepts utilisés sur 4 691 termes du dictionnaire. Contrairement aux concepts cellulaires, les termes cités dans le corpus sont hétérogènes car ils comportent des organes immunitaires (ex : LYMPH NODE) et des systèmes non immunitaires (ex : NERVOUS SYSTEM). Certains concepts ne sont pas informatifs (ex : ANATOMICAL SYSTEM). Le concept BLOOD est le plus cité dans plus de 40 000 abstracts (Figure 32).

- **Les concepts pathologiques**

  Le dictionnaire issu de l'onotologie *Human Disease Ontology 2015* est le dictionnaire le plus riche. Il comporte 22 957 termes, dont 3 695 ont été utilisés dans le corpus CVSI. La pathologie la plus citée est *CANCER* dans plus de 14 000 résumés. Néanmoins, les pathologies sont d'origines diverses. On retrouve les pathologies d'origines infectieuses (ex : TETANUS), les maladies dues à des réactions d'hypersensibilité (*ASTHMA*), mais les cancers sont les pathologies les plus citées (Figure 33).

Pour les gènes et leurs produits, l'utilisation du package GENIA tagger nous a permis d'annoter :

- 19 426 concepts protéiques cités
- 453 concepts ARN cités
- 2 234 concepts ADN cités

- Les protéines sont les concepts les plus cités : 19 426 concepts cités dans 56 562 sur les 120 000 résumés. Le concept le plus cité est *CYTOKINE*, un concept général que se réfère à une classe de molécules immunitaires. Néanmoins, les résultats d'annotation démontrent qu'une grande partie de cytokines spécifiques sont très représentées (*IL-2, IL-10*). La liste comprend aussi des antigènes (*CD4, CD8*). Les ARN sont les concepts les moins cités : 453 concepts dans 11 850 résumés sur les 120 000 résumés. Une partie des concepts est redondante entre les Protéines et les ARN tels que le concept *CYTOKINE*. Il y a 2 234 concepts ADN cités dans 20 132 sur les 120 000 résumés. Les concepts sont des classes de gènes comme *DETOXIFICATION GENE* (les gènes de détoxifications) cité dans 1 156 résumés, des fragments d'ADN (BETA 2-BINDING SITE) ou des noms de gènes (OLR1 GENE).

Pour être intégré dans la signature, il faut impérativement que le gène ou son produit soit annoté dans le même corpus avec une population cellulaire, un compartiment anatomique et une pathologie. Nous avons décidé de considérer uniquement les concepts protéiques et les noms des ARN vue la nature des données générées (données transcriptomiques), que le package *GENIA tagger* est imprécis dans la classification ARN/protéine, et que les concepts ADN sont majoritairement peu informatifs. Le rendement d'annotation des protéines et des ARN étant faible (peut de concepts reconnus par rapport aux concepts cellulaires et pathologiques), nous avons gardés, pour la suite de l'analyse, les signatures où il y a au moins une protéine ou un ARN cités.

Pour conclure l'analyse des concepts identifiés, montre la spécificité de notre requête vue qu'une grande partie est spécifique du système immunitaire. A noter ici la nécessité d'utiliser un outil de normalisation pour les noms de gènes et de leurs dérivés pour un confort d'utilisation.

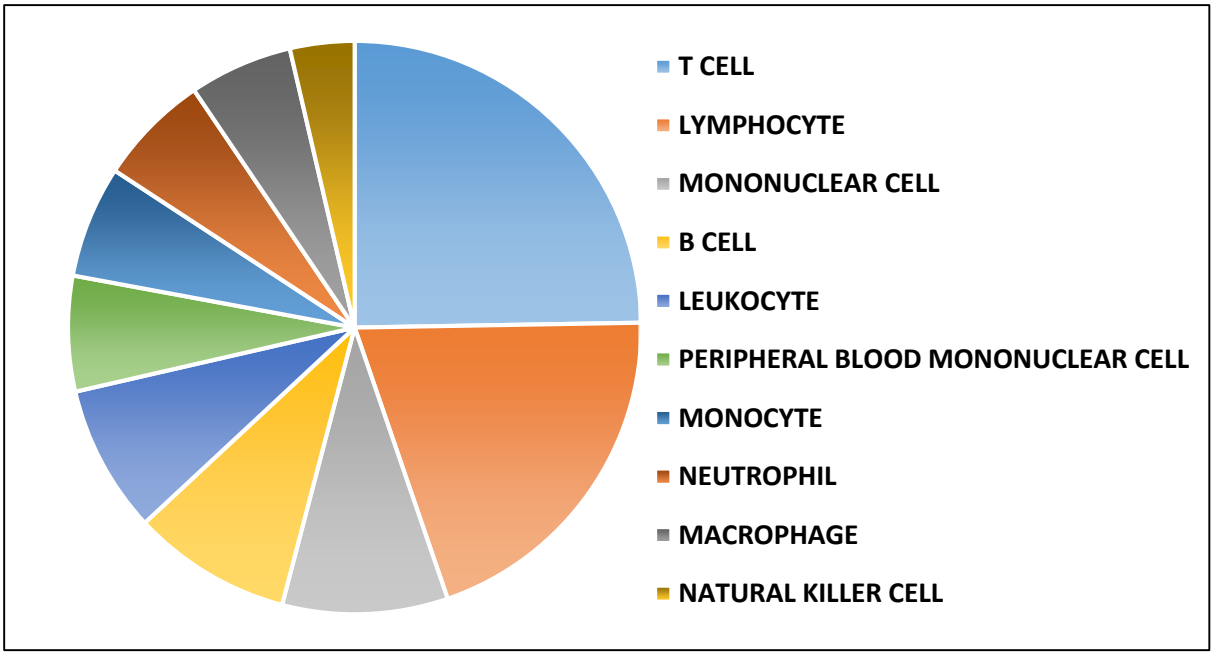


Figure 31. Graphique en camembert des 10 concepts cellulaires les plus cités dans le corpus sur le vieillissement du système immunitaire (CVSI).

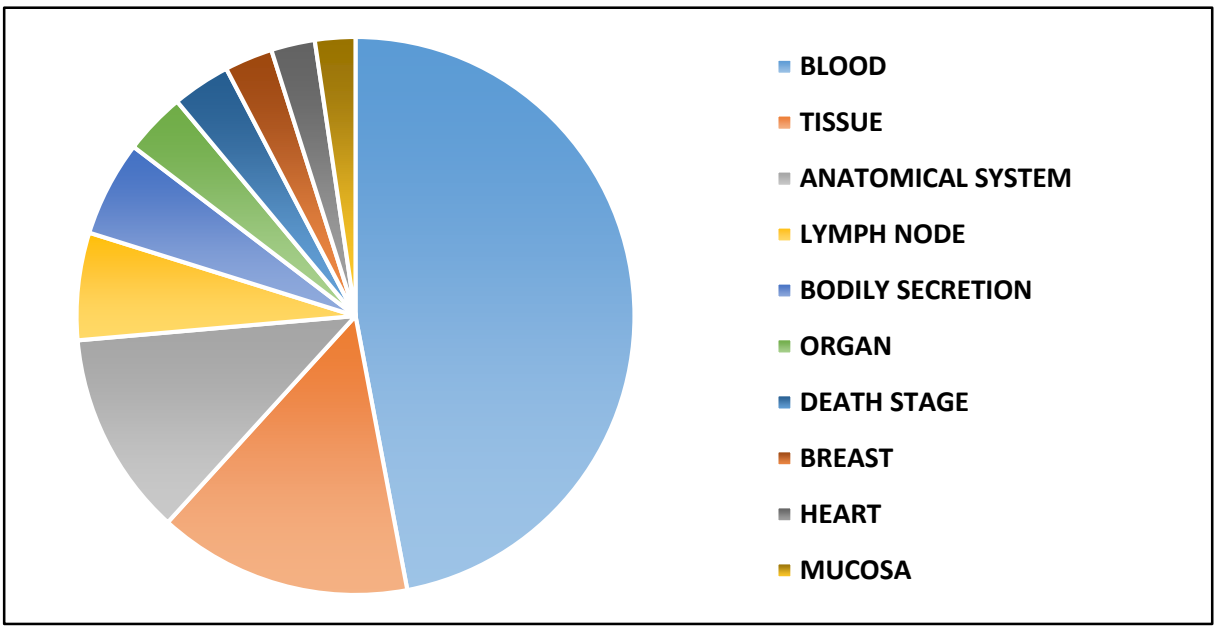


Figure 32. Graphique en camembert des 10 concepts anatomiques les plus cités dans le corpus sur le vieillissement du système immunitaire (CVSI).

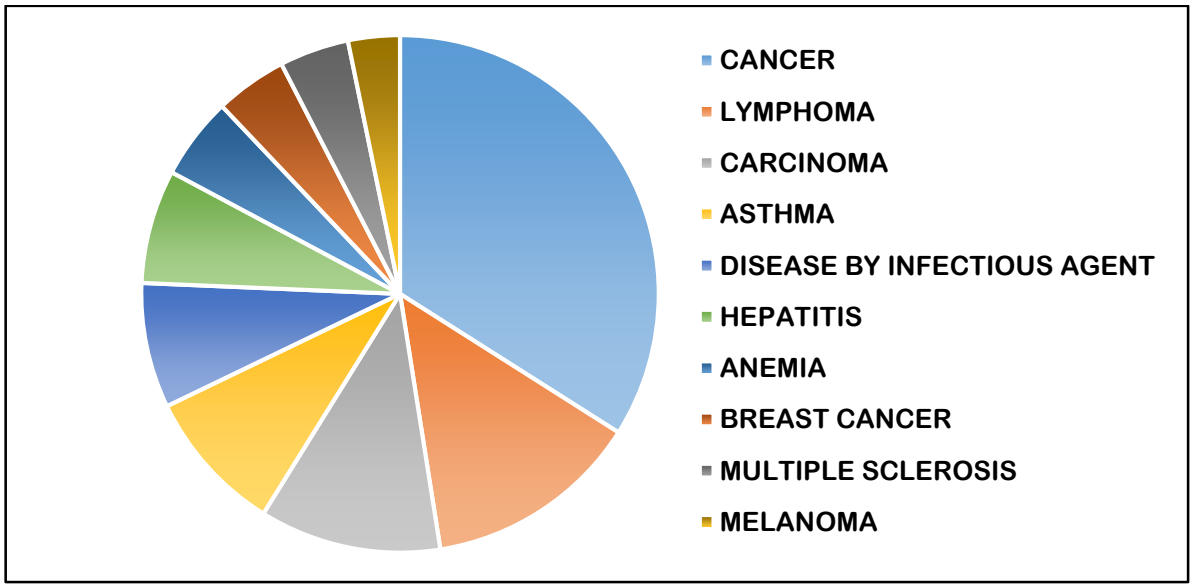


Figure 33. Graphique en camembert des 10 concepts pathologiques les plus cités dans le corpus (CVSI).

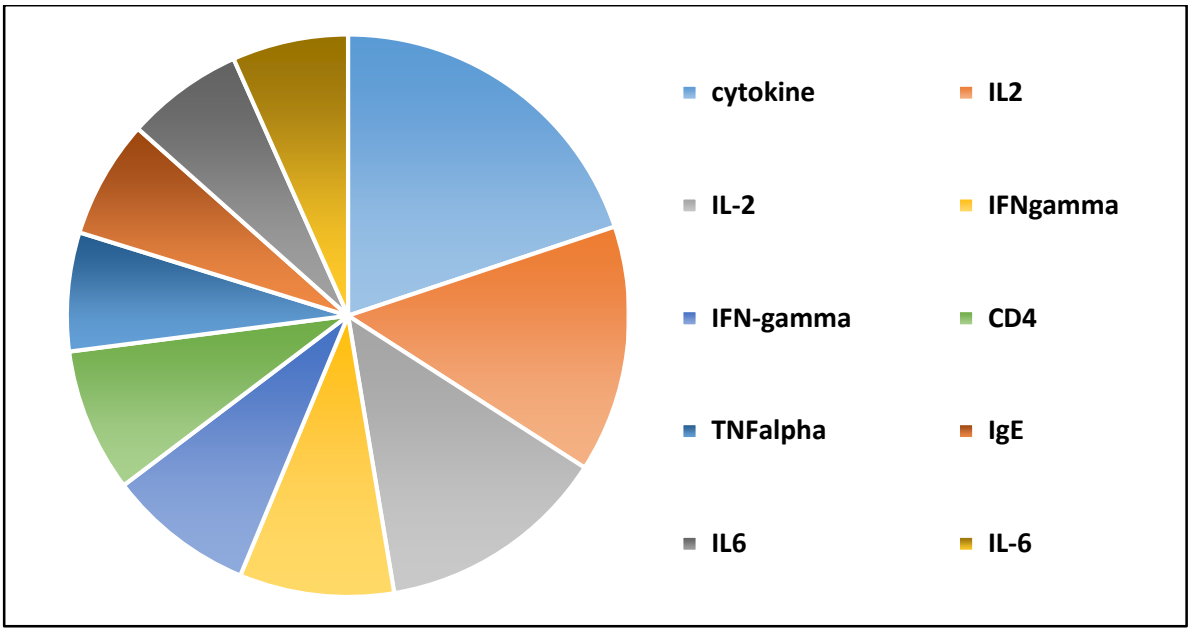


Figure 34. Graphique en camembert des 10 concepts protéiques les plus cités dans le corpus (CVSI).

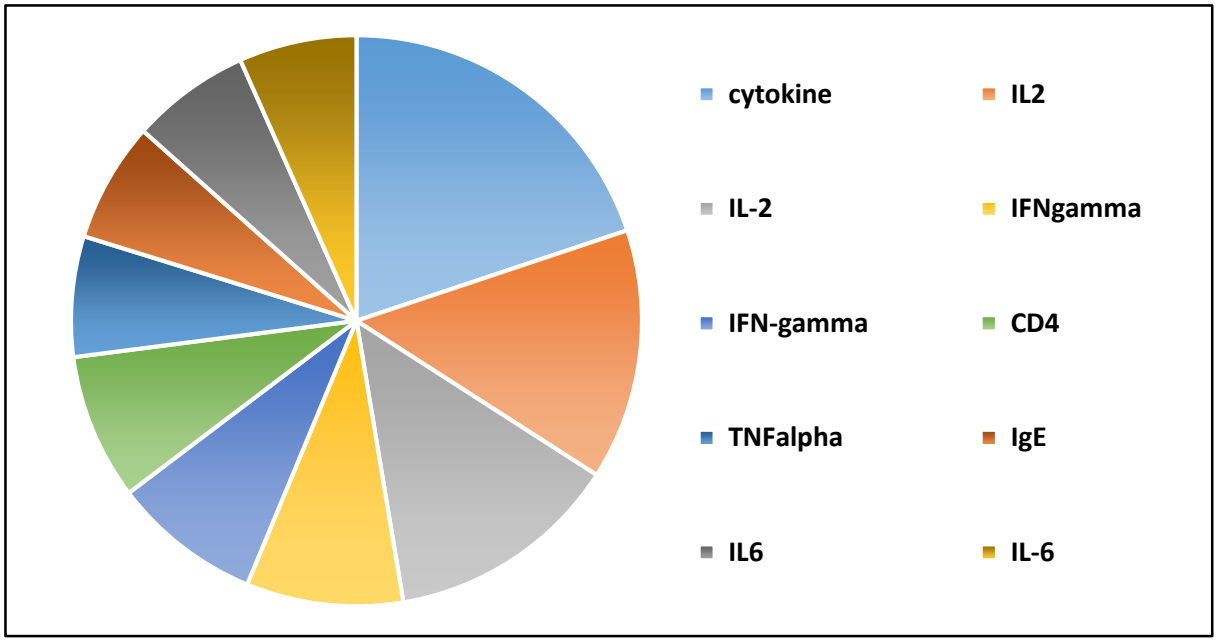


Figure 35. Graphique en camembert des 10 concepts ARN les plus cités dans le corpus (CVSI).

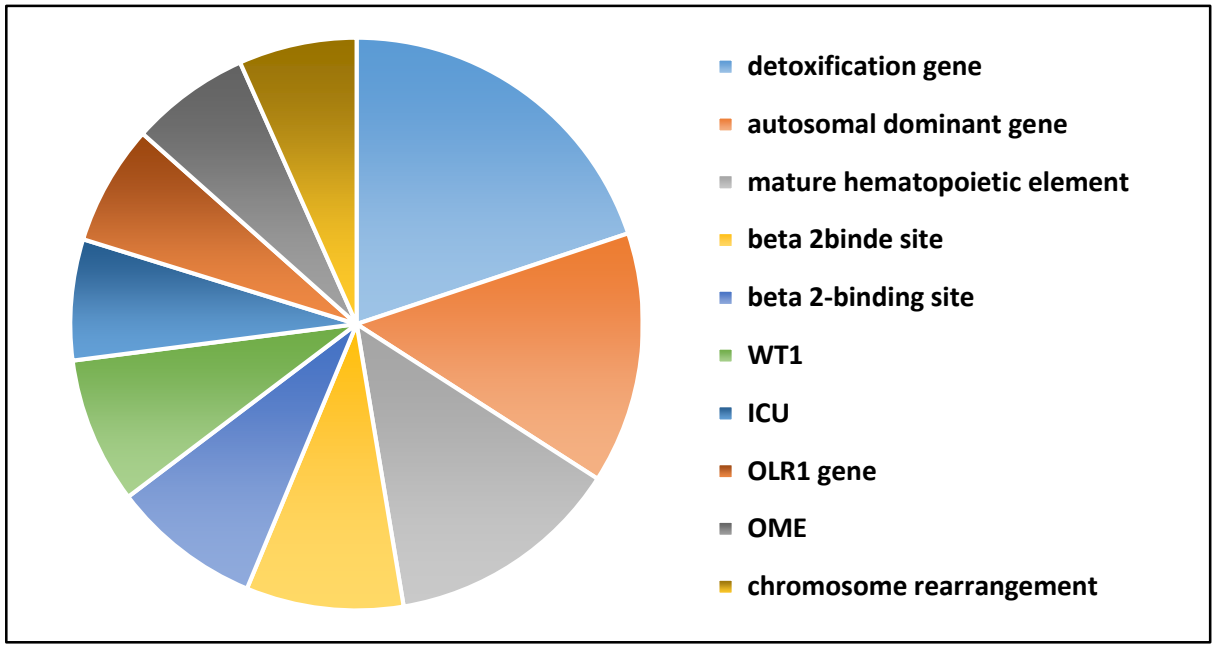


Figure 36. Graphique en camembert des 10 concepts ADN les plus cités dans le corpus (CVSI).

3. Comparaison des processus biologiques mis en jeux d'un contexte à un autre

Le contexte étant défini par les trois paramètres suivants : les populations cellulaires, la localisation anatomique et l'état pathologique, une modification d'un seul paramètre devrait être traduite par une modification et notre postulat de départ stipule que chaque contexte est caractérisé par une signature de gênes spécifique. Cette signature devrait être traduite par des processus biologique mises en jeux différents d'un contexte à un autre même si on fait varier une composante unique (soit la population cellulaire, soit la pathologie ou bien le compartiment anatomique) du contexte. Nous avons comparé les signatures de gènes et les processus biologiques mis en jeux issus de quatre contextes identifiés dans le corpus de vieillissement du système immunitaire. Les quatre contextes étudiés sont : T CELL\LYMPHNODE\CANCER, T CELL\BLOOD\CANCER, T CELL\BLOOD\DERMATITIS et BCELL\BLOOD\CANCER. Les signatures de gènes sont présentées dans l'annexe 6. Pour l'analyse des processus biologiques impliqués nous avons décidé d'utiliser les annotations de *l'ontologie Biological Process* de *Gene Ontology* et avons conçu l'expérience suivante :

- Nous avons procédé dans un premier temps à une annotation des signatures en utilisant l'interface Web de *Gene Ontology*.

- Nous avons utilisée l'outil PANTHER pour filtrer les annotations de l'ontologie *Biological Process.*
- Le résultat est présenté sous forme de camemberts qui regroupent les annotations par catégorie de processus (Voir Figure 37)
- Nous avons par la suite comparé les catégories par contexte

L'analyse du résultat montre qu'une partie des processus biologiques sont partagés par les quatre contextes. Il s'agit essentiellement des processus énergétiques (les processus métaboliques), les processus indispensables pour le cycle cellulaire (processus cellulaires, processus de développement…), les processus immunologiques -les populations cellulaires choisies sont des cellules immunitaires- (processus du système immunitaire, réponse aux stimuli). Néanmoins si on compare d'un contexte à un autre on peut identifier des processus spécifiques, tels que la croissance qui est spécifique d'un seul contexte (T CELL/LYMPH NODE/CANCER), la locomotion spécifique de deux contextes (T CELL/LYMPH NODE/CANCER et B CELL/BLOOD/CANCER) et la localisation et la reproduction qui sont spécifiques de trois contextes. Pour conclure on peut dire que chaque contexte possède sa propre signature de processus biologique mis en jeux.

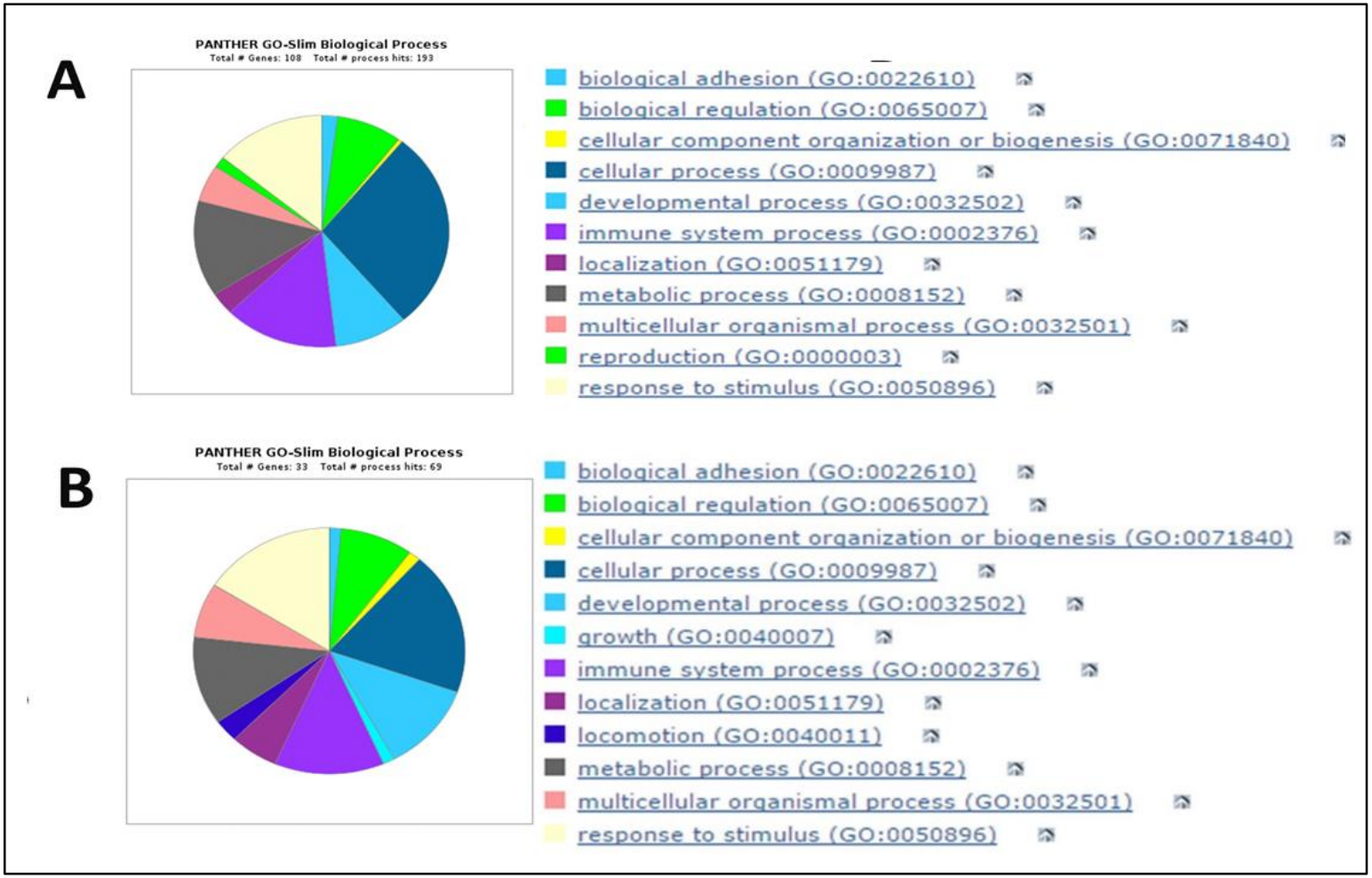


Figure 37 A/B. Résultat de l'annotation des signatures de gènes de quatre contextes identifiés à partir du corpus sur le vieillissement du système immunitaire. Cette annotation a été faite avec les concepts *Biological Process* de *Gene Ontolotgy*, en utilisant l'outil PANTHER. A : les gènes du contexte, T CELL/BLOOD/CANCER, B : les gènes du contexte, T CELL/LYMPH NODE/CANCER.

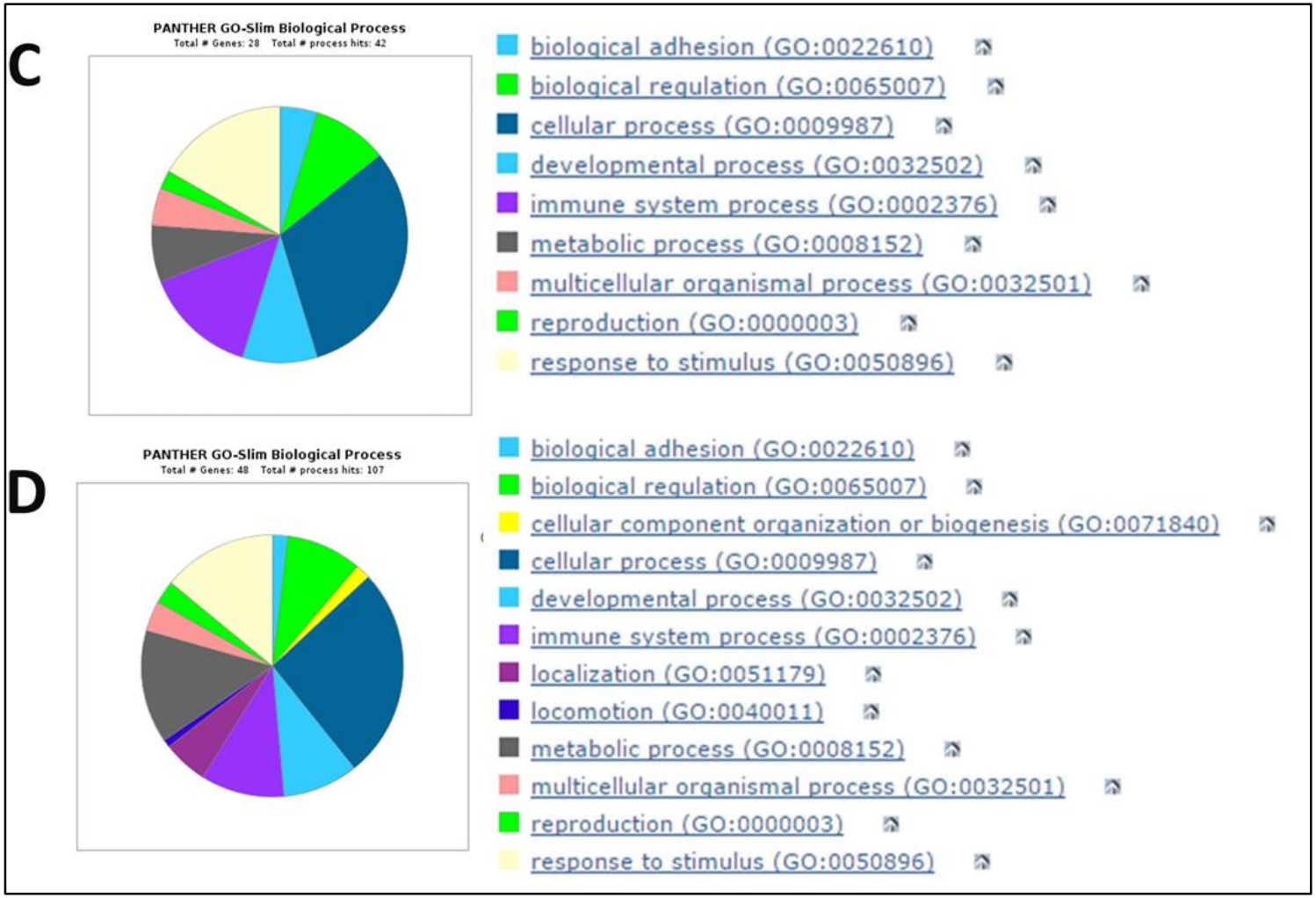


Figure 37 C/D. Résultat de l'annotation des signatures de gènes de quatre contextes identifiés à partir du corpus sur le vieillissement du système immunitaire. Cette annotation a été faite avec les concepts *Biological Process* de *Gene Ontolotgy*, en utilisant l'outil PANTHERC : les gènes du contexte, T CELL/BLOOD/DERMATITIS, D : B CELL/BLOOD/CANCER.

E. Discussion : OntoContext, contextualisation et théorie de l'information

1. OntoContext

Pour cette étape de contextualisation (nous discutons dans le paragraphe suivant du choix du contexte), nous avons développé une méthode d'annotation textuelle pour l'identification des concepts cellulaires, anatomiques et pathologiques co-cités avec des noms de gènes et/ou de leurs produits. OntoContext, utilise trois ontologies distinctes :

- *Cell Ontology*, pour les populations cellulaires (Bard et al. 2005)
- *UBERON Ontology*, pour les localisations anatomiques (Mungall et al. 2012)
- *Human Disease Ontology*, pour les concepts pathologiques (Kibbe et al. 2014)

et le package GENIA tagger pour les noms de gènes. OntoContext est fourni actuellement avec ces trois ontologies ; un point d'amélioration est l'intégration d'un module de prétraitement des ontologies pour permettre une souplesse d'utilisation. Ainsi l'utilisateur pourra personnaliser le processus d'annotation en fonction de ces besoins en utilisant les ontologies qui l'intéresse.

1.1. Choix des outils

Nous avons choisi d'utiliser trois ontologies différentes. Ceci nous permet d'identifier séparément chaque genre de concepts et d'éviter, par exemple, la confusion qui pourrait survenir entre compartiments anatomiques et populations cellulaires, ainsi qu'entre populations cellulaires et lignées cellulaires, comme c'est le cas pour les outils qui se basent sur des dictionnaires d'annotation issus des MeSH terms tels que l'outil ANNI (Jelier et al. 2008) et GoPubMED (Doms and Schroeder 2005). Nous avons utilisé OntoContext pour annoter un corpus de plus de 120 000 résumés. Nous avons trouvé plus de 2 313 concepts cellulaires (populations cellulaires) utilisés qui font plus de 300 000 annotations textuelles. Or, dans l'outil ANNI les articles pré-annotés de MEDLINE comme le confirme les auteurs, il y a seulement 1724 annotations cellulaires (population + culture). Ceci est dû à la richesse de l'ontologie Cell Ontology par rapport aux termes MeSH. Par exemple, le terme MeSH *Thymocyte* constitue une feuille de la hiérarchie (il n'est pas explicité) ; en revanche, dans la terminologie *Cell Ontology*, ce terme est bien plus détaillé avec beaucoup de termes fils. Les outils ANNI et GoPubMED sont des interfaces web et des applications exécutables ce qui rend leur utilisation limitée. C'est pourquoi, notre choix s'est orienté vers le développement d'un package Python qui permet une souplesse d'utilisation et d'intégration dans d'autres applications. De plus, Python est un

langage de programmation très partagé par la communauté des biologistes et des bioinformaticiens pour le développement d'applications. L'outil ScLite (Venkatesan et al. 2016) peut être interrogé via des API, mais ne permet pas d'identifier les populations cellulaires indispensables pour la définition du contexte. OntoContext permet l'annotation directe de corpus de textes permettant ainsi à l'utilisateur de choisir son propre corpus et de personnaliser la recherche en fonction de l'objectif. Ici, nous avons présenté son utilisation pour l'étude du vieillissement du système immunitaire en utilisant un corpus d'abstracts à partir de la base de données MEDLINE. On peut aussi l'utiliser pour cribler des textes entiers ; la seule condition c'est le format texte. L'outil ANNI, quant à lui, utilise des corpus de textes pré-annotés mais nous ne disposons d'aucune information sur le corpus utilisé, ni sur la mise à jour des bases d'annotation. L'outil NCBO Annotator[49] (Jonquet et al. 2009) du centre national des bio-ontologies de l'université de Stanford est un outil de référence qui permet cet exercice.

1.2. OntoContext versus NCBO Annotator

La comparaison entre NCBO Annotator (l'outil de référence pour l'annotation des textes) et le module Annot de notre package OntoContext, démontre de meilleures performances pour OntoContext en matière de rappel (29% contre 21% pour les pathologies respectivement pour OntoContext contre NCBO Annotator et 60% contre 23% pour les populations cellulaires). Cette mesure renseigne sur la perte de l'information (Sokolova et al. 2006). Pour les populations cellulaires par exemple, 40% des populations identifiées par les experts lors de l'annotation manuelle, ne l'ont pas été par l'annotation automatique par OntoContext contre 77% pour NCBO Annotator. Cette meilleure performance d'OntoContext est due en partie au prétraitement pour intégrer les pluriels ce qui n'est pas le cas de NCBO Annotator. Nous pouvons déduire l'importance du pluriel en comparant surtout les performances des populations cellulaires, souvent citées au pluriel dans les publications.

La précision est une mesure qui permet d'évaluer le taux de concepts faussement identifiés par les outils (Sokolova et al. 2006). Ces faux positifs peuvent être assimilés au bruit de fond. Pour cette mesure, les performances d'OntoContext et de NCBO Annotator sont proches (78% contre 81% pour les populations cellulaires, et 61% contre 59% pour les maladies respectivement pour OntoContext et NCBO Annotator). Cette notion de faux positifs (non identifiés par les experts) est sujette au débat. En effet dans les corpus de référence (Rosario and Hearst 2004, Bada et al.

[49] https://bioportal.bioontology.org/annotator

2012). On peut noter que la validation est l'identification manuelle des concepts par des experts qui représente un effort humain sujet à des erreurs et oublis : tous les concepts identifiés par l'annotation automatique (OntoContext ou NCBO Annotator) sont cités dans le texte. Pourtant, ils ne sont pas considérés comme tels car ils n'ont pas été reconnus par les annotateurs manuels. Ceci pause la problématique quant à la fiabilité des corpus de validations et surtout la référence à adopter la machine ou l'homme. Surtout quand on admet qu'OntoContext et NCBO Annotator ont des limites notamment pour la reconnaissance des concepts éclatés. Un exemple de concepts éclaté est CELLS IN ... EPITHELIAL ... REGIONS. Il est entrecoupé par d'autres mots (représenté par les trois points). Ce concept n'a été reconnu ni par OntoContext ni par NCBO Annotator. En effet, pour qu'il soit retenu lors de l'annotation automatique, il faut que tous les mots composant ce concept soit cités les uns derrières les autres d'une matière continue (ex : CELLS IN EPITHELIAL REGIONS). C'est un point qui sera abordé pour l'amélioration d'OntoContext.

1.3. Etude du vieillissement du système immunitaire

Nous avons choisi d'étudier un corpus de résumés d'articles sur le vieillissement du système immunitaire à partir de la base bibliographique MEDLINE. Les résumés des articles sont disponibles gratuitement contrairement aux articles complets (les articles complets sont souvent payants et des restrictions peuvent exister pour le téléchargement) et facilement téléchargeable via PubMed sous format texte.

La requête permettant d'identifier et extraire les résumés des publications qui traitent du vieillissement du système immunitaire en utilisant les termes MeSH de MEDLINE est la suivante :

> *"((aging OR age groups) AND (Proteins OR genes OR antigens) AND (human OR mice) AND immune system)"*

Nous avons utilisé le système MeSH pour la requête, pour nous assurer de la pertinence des résumés d'articles identifiés. La recherche bibliographique est en effet une première étape de contextualisation. Cependant, les publications dans MEDLINE sont diverses et traitent de tous les sujets qui touchent à la recherche biomédicale. Par conséquent, l'étape de recherche bibliographique est une étape cruciale, une recherche erronée pouvant introduire des erreurs dans l'interprétation des résultats. Nous avons ainsi décidé de filtrer les résumés à partir des 1980, date du début de l'immunologie moderne (Doherty and Robertson 2004). En effet, l'optimisation des techniques de cytométrie en flux et des techniques de PCR y ont beaucoup

contribué (Ermann et al. 2015) pour le phénotypage des cellules et des molécules immunitaires. Cette étape de filtrage nous a permis de réduire le nombre de résumés à 123 393. L'analyse des concepts utilisés pour l'annotation des populations cellulaires montre que la plus grande partie des populations identifiées consiste en des populations immunitaires (sur les 50 concepts cellulaires les plus cités, un seul n'est pas immunitaire « *MUSCLE CELL* » et le concept « *CELL* » est un concept général). Ceci est dû à la nature de la requête que nous avons lancée dans PubMed. Ceci est un indicateur aussi de la qualité de la requête. Pour les concepts anatomiques et pathologiques, le résultat est plus hétérogène. En effet, le rôle du système immunitaire étant la préservation de l'intégrité de l'organisme, il interagit via ses cellules et ses molécules avec tous les organes du corps qui interviennent dans tous les processus pathologiques qui l'affectent. Néanmoins, on note ici une prédominance des concepts pathologiques du cancer (17 concepts sur les 50 les plus cités ; le concept « Cancer » est le concept le plus cité, plus de 14 000 fois). Cette prédominance traduit une tendance des sujets d'étude sur le cancer pour notre corpus. En effet, des études montrent une augmentation de l'incidence du cancer chez les personnes âgées (en France selon l'Institut national du cancer l'âge moyen de diagnostic est de 68 ans[50]). Sur un plan physiologique, des liens ont été démontrés entre les gènes responsables de la correction de l'altération de l'ADN (l'altération de l'ADN est une des causes du cancer) et le ralentissement des processus de vieillissement (Hasty et al. 2003).

Pour les concepts pathologiques, anatomiques et cellulaires, l'utilisation des relations de synonymies nous permet de réduire le nombre de concepts identifiés en regroupant les concepts synonymes ; c'est ce qu'on appelle normalisation. Par exemple, les textes qui contiennent les concepts « *T CELL* », « *T-CELL* », « *T-LYMPHOCYTE* » et « *T LYMPHOCYTE* » seront regroupés sous le concept unique « *T CELL* ». C'est ce qui explique aussi que seulement 2 313 concepts ont été mentionnés 300 000 fois. Dans cette étape de normalisation, l'agrégation des synonymes limite aussi la dispersion de l'information. Par exemple, si on veut étudier les gènes du contexte « *T CELL* », « *BREAST* », « *BREAST CANCER* », OntoContext permet de regrouper et entrecroiser tous les textes dans lesquels ces termes sont cités (Table 4).

[50] http://www.e-cancer.fr/Professionnels-de-sante/Les-chiffres-du-cancer-en-France/Epidemiologie-des-cancers

Table 4. Table concept-synonymes pour le concept *T CELL, BREAST, BREAST CANCER*

Concept	Synonymes
T CELL	*T CELL, T-CELL, T-LYMPHOCYTE* et *T LYMPHOCYTE*
BREAST	*BREAST, MAMMA, MAMMARY PART OF CHEST* et *MAMMARY REGION*
BREAST CANCER	*BREAST CANCER, BREAST TUMOR, MALIGNANT NEOPLASM OF BREAST, MALIGNANT TUMOR OF THE BREAST, MAMMARY CANCER, MAMMARY NEOPLASM, MAMMARY TUMOR* et *PRIMARY BREAST CANCER*

Cette perte d'information est limitée par l'utilisation du module *criscross* du package OntoContext qui permet aussi l'agrégation des termes fils. En revanche, pour l'identification des noms de gènes et de leurs produits nous avons intégrés le package *GENIA tagger* dans OntoContext. *GENIA tagger* ne permet pas de regrouper les termes synonymes. Ainsi, dans la présentation des résultats d'annotation, on peut lire « *IL2* » et « *IL-2* » qui sont deux synonymes non regroupés. Le non regroupement des synonymes explique que 19 424 concepts ont été utilisés pour seulement 56 562 annotations protéiques. A ce point de l'analyse, regrouper les noms de gènes et leurs en synonymes n'est pas indispensable vues que dans la deuxième partie nous utilisant deux bases de données (*db-string et GO annotation*) reconnaissant les différentes nomenclatures. Néanmoins, une approche de normalisation (agrégation par synonymie) devrait être intégrée à notre package.

Pour conclure, dans cette première partie de la discussion, abordé l'outil technologique de contextualisation que nous avons développé. Dans ce qui suit, nous allons nous intéresser à la justification de la contextualisation.

2. Contextualisation et théorie de l'information

La vision classique de la génétique voudrait qu'il y ait une relation causale entre un gène bien déterminé et un trait de caractère (Orgogozo et al. 2015). Cette vision réductionniste de la relation causale « gène/phénotype » a été remise en cause notamment par les spécialistes de la théorie de l'information. Un des arguments remis en question par les théoriciens de l'information est le caractère déterministe de l'information génétique (causalité gène/phénotype ≈ gène/fonction). Elle serait une relation stochastique et non pas déterministe. Selon Longo et coll. (Longo et al. 2012) :

« *Un phénotype est déterminé par un ou plusieurs gènes et un gène peut être responsable d'un ou de plusieurs phénotypes* ».

Or, depuis l'expérience de Jacob et Monod sur le promoteur LacZ (Jacob and Monod 1961), nous savons que l'expression des gènes dépend aussi des conditions environnantes : c'est la présence du lactose dans le milieu qui conditionne la transcription du gène LacZ. Dans la continuité, la biologie développementale propose une vision plus réaliste de la relation gène/fonction. Cette discipline présente plutôt cette relation comme suit (Orgogozo et al. 2015) :

$$\left\langle \begin{array}{c} Gène(s) \\ Environnement\ cellulaire \\ Epigénétique \\ Environnement\ extracellulaire \\ Expérience\ individuelle\ \ldots \end{array} \right\rangle \rightarrow Fonction\ (ou\ phénotype)$$

Nous avons traduit cette relation par la contextualisation. En effet, nous partons de l'hypothèse suivante : un gène est exprimé différemment en fonction de la population cellulaire (Zhang et al. 1997, Palmer et al. 2006, Li et al. 2016), de la localisation anatomique (organe, tissu) (Nishimura et al. 2004, Wong et al. 2016) et en fonction de l'état pathologique (Yang et al. 2014, Calon et al. 2015). Chaque gène étant impliqué dans un ou plusieurs processus biologiques, qui assurent une ou plusieurs fonctions biologiques donc ces processus mis en jeux devraient être différents d'un contexte à un autre. Pour la confirmation de cette hypothèse, nous avons confronté les signatures de gènes issues de contextes proches et nous avons comparé les annotations aux données de *Biological Process* de *Gene Ontology*. Comme présenté dans la Figure 37, une partie des processus mis en jeux est partagé entre tous les contextes. En note que ces processus sont indispensable au fonctionnement cellulaires. D'un point de vue évolutif les processus métaboliques et les processus cellulaires sont des processus qui ont été conservé chez les organismes procaryotes et eucaryotes (Jensen et al. 2006, Peregrín-Alvarez et al. 2009). Ceci plaide en la faveur de la transmission et la transmission des fonctions biologiques durant l'évolution. Les gènes représentent leurs supports et donc le support de l'information biologique et cette information dépend du contexte d'expression puisque d'un contexte à un autre les processus varient. Les processus immunologiques sont des processus acquis avec les organismes pluricellulaires (Flajnik and Kasahara 2010) néanmoins le fait que les populations cellulaires des contextes choisis soit des cellules immunitaire explique que ces processus soit partagés. Selon le contexte étudié il y a des processus biologiques spécifiques mis en jeux. Ces

processus biologiques résultent de l'expression de fonction biologiques potées par les gènes, d'où l'importance du contexte dans l'étude et l'interprétation des résultats d'expression des gènes. En effet comme en la vue au début du chapitre le processus d'analyse et d'interprétation des résultats passe par une étape d'analyse statistiques et de regroupement pour la quantification, une étape d'annotation expérimentale pour étiqueter les résultats pour permettre leurs interprétations et tirer des conclusions. Notre interventions vient dans l'étape d'annotation expérimentale nous proposons ici un étiquetage en fonction des contextes d'expression ce qui permet d'apporter un complément d'information. L'analyse de la littérature permet de produire une base de contextes personnalisables selon le sujet d'étude. Ici, nous nous sommes intéressés à l'exemple du vieillissement du système immunitaire. Une des problématiques de l'expérience de départ est que les données produites l'ont été sur toutes les sous-populations de « *T CELL* » du « *THYMUS* ». Or, nous avons besoin d'identifier les taux d'expression des signatures génétiques de chaque sous-population des cellules T tel que « *HELPER T CELL* » et les « *REGULATORY T CELL* ». L'exploitation de la base de contextes permet d'identifier une liste de gènes pour chaque sous-population qui pourra être confrontée aux données biologiques dans une étape suivante. Le laboratoire LGIPH, quant à lui, utilise une approche expérimentale basée sur la PCR, utilisant un nombre réduit de gènes par expérience. Les expérimentateurs ont donc besoin d'identifier des cibles. L'approche de contextualisation permet aussi le criblage de cibles génétiques pour les études. Plus généralement, l'étude du contexte d'expression aide à comprendre la relation gène > fonction comme le préconise les spécialistes de la biologie fonctionnelle. Néanmoins, pour assurer une fonction biologique, plusieurs gènes interagissent. Pour élucider les interactions génétiques qui assurent ces fonctions, nous présentons dans le chapitre suivant une approche de modélisation des voies biologiques.

Chapitre 3 : Reconstitution des Voies Biologiques (Contribution)

A. Introduction et objectifs de la reconstitution des voies biologiques

Dans le chapitre 2, nous avons abordé la question de la contextualisation des gènes et de leurs produits et nous avons présenté l'outil OntoContext, que nous avons développé pour gérer des listes de gènes regroupés par contexte. Nous avons vu que les profils d'expression peuvent différer d'un contexte biologique à un autre. Il faut alors élucider les mécanismes moléculaires qui y interviennent. Le but de cette étape est de générer une base de connaissances regroupant des réseaux de gènes impliqués dans les mêmes voies biologiques automatiquement. Pour cela, nous avons utilisé deux approches : une approche basée sur l'utilisation des distances sémantiques et une seconde approche basée sur l'utilisation d'un algorithme génétique.

Pour aborder cette partie, il est important de définir un concept clé : le concept de voie biologique. Le NIH (National Institute of Health) propose la définition suivante d'une voie biologique :

> *« C'est une série d'actions, impliquant plusieurs molécules dans une cellule, qui conduisent à un certain produit ou à un changement dans une cellule. Une telle voie peut déclencher l'assemblage de nouvelles molécules, telles qu'une graisse ou une protéine. Les voies peuvent également activer ou désactiver des gènes et stimuler une cellule à se multiplier, à se différencier, à s'autodétruire, à phagocyter, à se déplacer*[51]*… »*

Il découle de cette définition une classification des voies biologiques en trois types :

1) Les voies métaboliques : ensemble de réactions biosynthétiques et hydrolytiques, faisant intervenir des enzymes pour la synthèse et l'hydrolyse des composants biochimiques indispensables au fonctionnement cellulaire.
2) Les réseaux de régulation génétique : ensemble des mécanismes de régulation mis en œuvre pour passer de l'information génétique incluse dans une séquence d'ADN à un produit de gène fonctionnel (ARN ou protéine).
3) Voies de signalisation : système complexe qui assure la communication intra- et inter-cellulaire pour coordonner leurs activités ou bien assurer un ensemble de processus fondamentaux.

[51] https://www.genome.gov/27530687/

Cette classification paraît très synthétique mais d'autres définitions ont été avancées. *Gene Ontology*, une méta-ontologie contenant une ontologie de processus biologiques[52] (*biological process*), définit un processus biologique comme un ensemble d'événements assurés par une ou plusieurs fonctions moléculaires d'une manière organisée. KEGG[53], en revanche, propose la définition suivante : un réseau biologique est une représentation de l'état de nos connaissances sur les interactions moléculaires et les réseaux de réactions biologiques.

Nous retenons la définition de voie biologique comme un ensemble de gènes ou de produits de gènes qui interagissent pour assurer une fonction biologique. De ce fait, nous émettons l'hypothèse que des gènes ou des produits de gènes avec des annotations en « *biological process* » proches seraient impliqués dans les mêmes voies biologiques. D'autres auteurs ont émis cette hypothèse (Guo et al. 2005, Pesquita et al. 2009). Dans ce contexte, nous allons présenter deux approches basées sur la proximité sémantique calculée sur la base des annotations « *biological process* » de *Gene Ontology*. La première utilise uniquement les similarités sémantiques pour le regroupement des gènes. La seconde approche est développée en collaboration avec le Laboratoire LIPAH qui utilise la proximité sémantique et les données d'interactions. Pour la validation de ces approches nous nous sommes affranchies de l'exemple du vieillissement du système immunitaire. Le but étant de développer une approche généraliste. Dans ce chapitre nous faisons aussi appel à des notions sur la théorie des graphes. Historiquement, la première problématique scientifique liée au graphe a été posée par Leonhard Euler en 1736 dans son problème des sept ponts de la ville de Königsberg (Sandifer 2007). Le problème posé par Euler consiste à voir si, à partir d'un point donné, on peut passer par tous les ponts de la ville une seule fois (Figure 38). C'est à partir de ce problème que va émerger la théorie des graphes. Ainsi, on définit un graphe comme un ensemble de points appelés aussi nœuds ou sommet reliés entre eux par des traits, ou flèches qu'on appelle aussi arrêtes (Alba 1973). On parle ainsi de graphe orienté si les arrêtes ont un sens d'un sommet A vers un sommet B avec une relation asymétrique entre les sommets. Dans le cas contraire, si les relations sont symétriques et n'ont pas de sens on parle de graphe non orienté.

[52] http://geneontology.org/page/ontology-documentation

[53] http://www.genome.jp/kegg/pathway.html

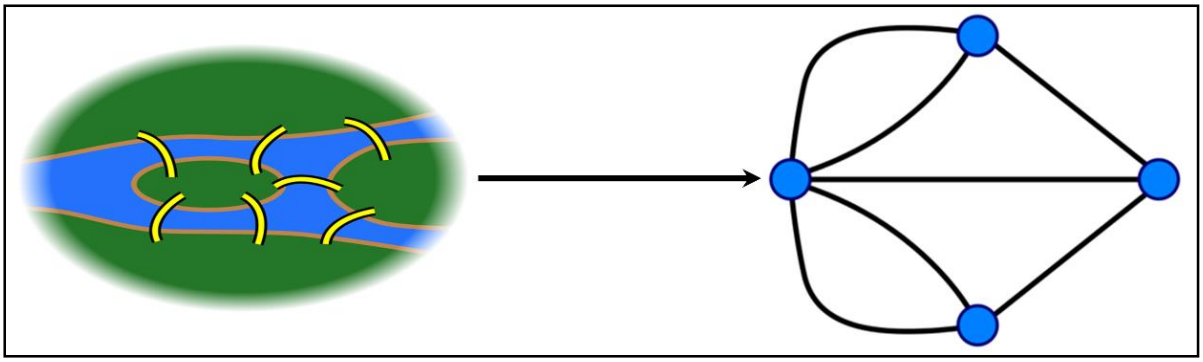


Figure 38. Modélisation du problème d'Euler. A gauche, une représentation des sept ponts de la ville de Königsberg. A droite la modélisation du problème sous forme de graphe : les arrêtes représentent les ponts, et les nœuds représentent les régions de la rivière reliant les ponts.

Plusieurs topologies de graphes ont été décrites dans la littérature, elles sont présentées dans la Figure 39 :

- Les graphes structurés :
  - Graphes homogènes où les sommets et les arrêtes décrivent une structure régulière comme la structure du filet de pêcheur.
  - Graphes hiérarchiques où les sommets et les arrêtes décrivent une structure pyramidale.
  - Graphes cycliques où les sommets et les arrêtes peuvent s'organiser en cycle comme les graphes circulaires ou en filet de pêcheur. Contrairement aux graphes acycliques qui ne permettent pas de revenir à un point départ.
  - Graphes polaires où tous les sommets sont rattachés à un seul point, on distingue aussi les graphes multipolaires où quelques points concentrent la plus grande partie des liens.
- On peut aussi citer les graphes quelconques, ce sont des graphes qui n'ont aucune particularité topologique.

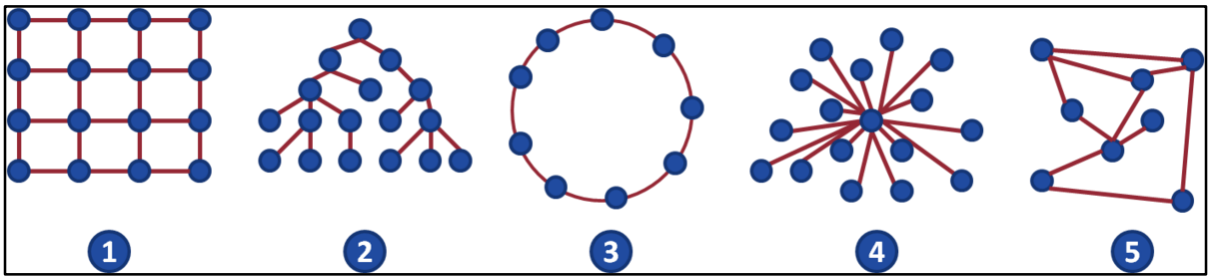


Figure 39. Schémas récapitulatifs des différentes topologies des graphes. 1. Graphe structuré dit en filet de poisson. 2. Graphe hiérarchique. 2. Graphe circulaire. 4. Graphe polaire. 5. Graphe quelconque. Sources : Par Goel — Travail personnel, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=48274506.

Dans ce qui suit il sera fait appel à cette notion de graphe surtout pour les ontologies qui constituent des graphes acycliques orientés (*Direct Acyclic Graph* DAG).

B. Gene Ontology, distance sémantique et modélisation des voies biologiques

Nous détaillons ici une première piste pour la modélisation des réseaux biologiques basée sur la proximité sémantique des annotations *Biological Process* de GO. Nous avons décidé de détailler la présentation de *Gene Ontology* car c'est là notre source centrale de données.

1. Gene Ontology et distances sémantiques

Comme nous l'avons mentionné dans l'introduction, GO est le fruit d'une collaboration au sein d'un consortium constitué au début par :

- les développeurs de la FlyBase[54] (Information et al. 1999), une base de données de génétique et biologie moléculaire pour la drosophile,
- Mouse Genome Informatics[55] (Eppig et al. 2017), la base internationale d'informations sur les souris de laboratoire,
- et *Saccharomyces* Genome Database[56] (Cherry 2015), la base d'information sur le *Saccharomyces cerevisiae*.

Ainsi, GO est un vocabulaire structuré, contrôlé, commun et défini avec précision pour décrire les rôles des gènes et des produits géniques dans n'importe quel organisme (Ashburner et al. 2000). Ce vocabulaire a été hiérarchisé en trois ontologies.

- ✓ *Biological Process* : Les processus biologiques peuvent être définis comme l'ensemble des objectifs biologiques auxquels les gènes ou leurs produits contribuent. Les processus impliquent souvent une transformation chimique ou physique (Ashburner et al. 2000). On peut citer l'exemple du concept général « *cell growth and maintenance* » ou bien d'un processus plus spécifique « *pyrimidine metabolism* ».
- ✓ *Molecular Function* : Il s'agit de l'ensemble des concepts qui décrivent l'activité biochimique des gènes et de leurs produits. Cette description se limite seulement à

[54] http://flybase.org

[55] http://www.informatics.jax.org/

[56] http://www.yeastgenome.org/

l'activité sans notion de localisation (anatomique ou subcellulaire). On peut citer comme exemple les concepts généraux « *enzyme* », « *ligand* » ou plus spécifiques « *adenylate cyclase* ».

- ✓ *Cellular component* : Cette ontologie se réfère aux emplacements subcellulaires où les produits des gènes sont actifs. A noter ici que les concepts concernent seulement l'état des connaissances sur les cellules eucaryotes. On peut citer comme exemple les concepts « *proteasome* » ou bien encore « *Golgi apparatus* »

Ce vocabulaire organisé de façon hiérarchique constitue une base d'annotation (attributs) pour les gènes, leurs produits et les complexes de produits de gènes. L'organisation en graphes acycliques orientés (hiérarchie + relations) de ces annotations, nous permet de calculer une proximité entre les différents gènes. C'est ainsi qu'ont été développées les distances sémantiques. Notre hypothèse de travail dans ce contexte est la suivante :

Les gènes qui sont proches sémantiquement pourraient être impliqués dans les mêmes processus biologiques.

1.1. Les bases d'annotation de Gene Ontology

Sur le site de *Gene Ontology*[57], les bases d'annotations qui relient les concepts (vocabulaire) aux produits de gènes sont classés par espèce. Pour l'espèce humaine (*Homo sapiens*), trois bases de données d'annotation (GOA) sont maintenues par le projet UniProt-GOA sous l'égide de l'European *Bioinformatics Institute* (après la mise à jour du 11/02/2016) :

- *Homo sapiens GOA* : contient toutes les annotations GO pour les références UniProt spécifiques de l'espèce humaine (19 230 produits annotés par 39 2941 concepts)
- *Homo sapiens (complex) GOA* : pour l'annotation des complexes macromoléculaires comporte 156 annotations pour 74 produits.
- *Homo sapiens (rnas) GOA* : pour l'annotation des ARNs, comporte 2017 annotations pour 156 produits.

1.2. Similarité sémantiques

En général, la similarité sémantique est une notion utilisée pour mesurer la proximité entre deux concepts appartenant à la même ontologie ou à deux ontologies différentes. Le

[57] http://www. geneontology.org

développement de *Gene Ontology* a permis l'émergence d'une nouvelle notion, la similarité fonctionnelle (Wang et al. 2007, Pesquita et al. 2009, Ruths et al. 2009). Cette notion est utilisée pour mesurer la proximité fonctionnelle entre deux produits de gènes en se basant sur la similarité sémantique entre leurs annotations fonctionnelles. Dans ce cadre, plusieurs mesures ont été développées dans cette partie nous nous sommes intéressé aux mesures entre termes dans l'annexe 7. II nous présentons quelques exemples d'extrapolation aux mesures entre gènes.

Gene Ontology est organisée en Graphe Acyclique Orienté (DAG). Les concepts (termes) sont des nœuds reliés par des arêtes orientées (Figure 40, le DAG présenté dans cette figure sera pris comme exemple pour toutes les distances).

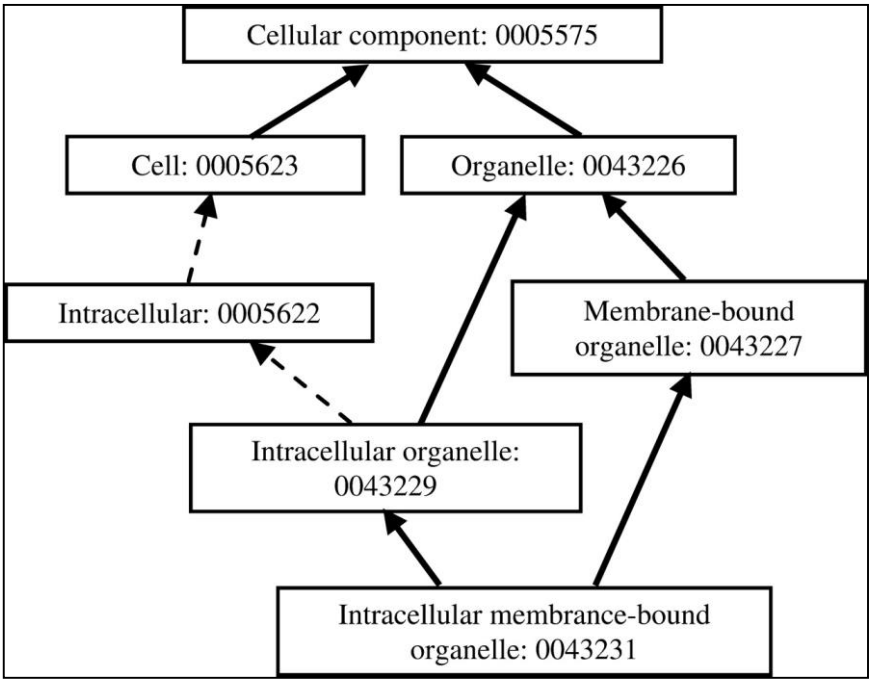


Figure 40. Un GO DAG extrait pour le terme « *Intracellular Membrane-bound Organelle :* 0043231 ». La racine de l'ontologie est le concept « Cellular componement » et la feuille est le terme « Intracellular Membrane-bound Organelle ». Ici les flèches en pointillés symbolisent les relations « Part_of » et les flèches pleines symbolisent les « Is_a » (Wang et al. 2007).

De cette organisation découle deux types de mesures :

- **Les mesures basées sur les arêtes** : ces mesures sont tout simplement basées sur le nombre d'arêtes entre deux nœuds (Pesquita et al. 2009) les plus connues sont :

  - La mesure de Wu & Palmer (Wu and Palmer 1994) :

$$Sim(a,b) = \frac{2 * \Delta(racine, c)}{\Delta(c,a) + \Delta(c,b) + 2 * \Delta(racine, c)}$$

Où : Δ = le nombre de nœuds du chemin entre deux concepts de l'ontologie (a et b)

c = l'ancêtre commun le plus bas (lcs) entre a et b

racine = la racine de l'ontologie

Donc si on veut calculer la similarité entre le terme « *Membrane-bound organelle : 0043227*» et le terme « *Intracellular organelle : 0043229*» :

$$Sim(0043227,0043229) = \frac{2*\Delta(0005575,0043226)}{\Delta(0043226,0043227)+\Delta(0043226,0043229)+2*\Delta(0005575,0043226)}$$

$$Sim(0043227,0043229) = \frac{2}{4} = 0{,}5$$

L'utilisation de cette distance pose plusieurs problèmes notamment celui des chemins à considérer : est-ce le chemin le plus long ? Le chemin le plus court ? Ou bien un chemin intermédiaire ? Cette question posée par plusieurs auteurs (Pesquita et al. 2009, Gan et al. 2013, Bettembourg et al. 2014) reste légitime si l'on considère que Wu et Palmer appliquent leur distance sur des ontologies simples avec un seul chemin possible entre les différents nœuds. Ce qui n'est pas le cas d'une ontologie aussi complexe que GO qui comporte plus de 45 000 termes[58] et 8 types de relations. Sur cette base, plusieurs distances ont été développées ou adaptées spécifiquement pour *Gene Ontology* tout en tenant compte de l'évolution de cette ontologie.

- La mesure de Pekar & Staab (Pekar and Staab 2002) :

$$Sim(a,b) = \frac{\delta(racine,c)}{\delta(a,c)+\delta(b,c)+\delta(racine,c)}$$

Ou : c = lcs entre a et b

δ = le plus court chemin

racine = la racine de l'ontologie

Pekar et Staab ont introduit dans leur approche la notion de similarité taxonomique donnée par la formule ci-dessus. Donc si on veut calculer la similarité entre le terme « *Membrane-bound organelle : 0043227*» et le terme « *Intracellular organelle : 0043229*» :

$$Sim(0043227,0043229) = \frac{\delta(0005575,0043226)}{\delta(0043226,0043227)+\delta(0043226,0043229)+\delta(0005575,0043226)}$$

[58] https://www.ebi.ac.uk/QuickGO/Dataset.html

$$Sim(0043227,0043229) = \frac{1}{3} = 0{,}33$$

- La mesure de Yu et coll. (Yu et al. 2005) :

$$\text{Sim(a,b)} = \begin{cases} \dfrac{\delta(\text{racine,a})}{\delta(\text{racine,b})}, \text{ si a et b sur la même branche et a plus proche de la racine} \\ \dfrac{\delta(\text{racine,b})}{\delta(\text{racine,a})}, \text{si a et b sur la même branche et b plus proche de la racine} \\ 0, \text{ si non} \end{cases}$$

A noter ici que contrairement à la première distance, Yu considère le plus long chemin entre deux nœuds notés ici δ. Donc si on veut calculer la similarité entre le terme « *Membrane-bound organelle : 0043227*» et le terme « *Intracellular organelle : 0043229*» :

$Sim(0043227,0043229) = 0$ *car les deux premières conditions ne sont pas vérifiés*

D'autres approches sont basées plutôt sur le nombre d'ancêtres en commun :

- La mesure de Wu et coll. (Wu et al. 2005) :

$$Sim\ (a,b) = \max_{La \epsilon Va, Lb \epsilon Vb} (Le\ nombre\ de\ termes\ en\ communs\ entre\ La\ et\ Lb)$$

Où V est le DAG représentant tous les chemins entre un terme et la racine. Donc si on veut calculer la similarité entre le terme « *Membrane-bound organelle : 0043227*» et le terme « *Intracellular organelle : 0043229*» :

$$Sim(0043227,0043229) = 2$$

Ces approches reposent sur deux hypothèses :

i. Les nœuds et les arêtes sont uniformément distribués sur la totalité de l'ontologie
ii. Les arêtes se trouvant sur un même niveau ont des distances équivalentes à celles entre termes du même niveau.

Or, ces conditions, et spécialement pour les ontologies biomédicales, ne sont pas vérifiées. Pour atténuer l'effet de cette hétérogénéité on peut à titre d'exemple ajouter un poids au nœud en fonction de la profondeur hiérarchique ou utiliser la densité des liens et considérer le type de lien (Pesquita et al. 2009).

- **Les mesures basées sur les nœuds** : dans une ontologie, particulièrement dans GO, un nœud représente un terme. Ce type de mesure est basé sur les propriétés intrinsèques des termes et des termes qui leur sont reliés (descendants ou parents). Ici une notion clé a été développée. C'est la notion du contenu en information (IC). Elle est similaire au contenu en information

développé par la théorie de l'information.Cette notion nous informe sur la spécificité et la normativité d'un terme (nœud) donné et est quantifiée pour un terme « c » donné de la manière suivante (Pesquita et al. 2009) :

$$\mathrm{IC} = -\log_2 p(c)$$

Où : p(c) = est la probabilité d'occurrence de « c » dans un corpus donné. Plusieurs auteurs considèrent p(c) comme la fréquence d'annotation d'un terme donné pour un gène donnée (Mazandu et Mulder 2014). C'est-à-dire que cette mesure est calculée sur la base du nombre d'occurrence d'un terme donné dans un ensemble de termes GO utilisés pour l'annotation d'un gène ou d'un ensemble de gènes. Pour cette raison nous n'allons pas produire un ensemble de calcul pour ces mesures.

Ici on peut citer la :

- La mesure de Resnik (Resnik 1995) :

  La similarité sémantique entre deux concepts a et b est :

$$sim\ (a,b) = (IC(MICA(a,b))$$

Il s'agit ici d'une application simple des définitions citées au-dessus. Si cette mesure permet de quantifier l'information partagée entre deux termes, elle ne permet, en revanche, pas d'intégrer les niveaux de précision de ces termes (la position des termes dans l'ontologie et par rapport au MICA (ancêtre commun le plus informatifs)).

- La mesure de Lin (Lin 1998) :

  La similarité sémantique entre deux concepts a et b est :

$$sim\ (a,b) = \frac{2 * IC(MICA(a,b))}{IC(a) + IC(b)}$$

Cette mesure présente l'inconvénient de ne pas tenir compte de la position dans l'ensemble de l'ontologie de ces concepts (a,b et MICA(a,b)). Par conséquent, des mesures entre des termes généraux peuvent être similaires à des concepts précis.

- **Les mesures hybrides :** Les approches essaient de combiner la quantification de l'information du terme et aussi sa position dans l'ontologie (son niveau de précision) pour pallier les limites des deux approches citées ci-dessus (basée sur les nœuds et sur les arêtes).

  - La mesure de Wang (Wang et al. 2007):

Pour mieux expliquer la distance, nous allons nous baser sur un exemple concret. Considérons le DAG (Graphe Acyclique Directe) GO présenté dans la Figure 40 et calculons

la contribution sémantique des parents dans le terme GO 0043231 donné par la formule suivante :

$$Sa(t)\begin{cases} Sa(a) = 1 \\ \max\{we * Sa(t') | t' \in termesfils(t)\} si\ t \neq a \end{cases}$$

avec *we*=poids des relations avec les termes fils (0,6 pour les relations *Part_of* et 0,8 pour les relations *Is_a)*

En nous basant sur cette formule, nous réalisons le calcul suivant :

- $S_{0043231}(0043231) = 1$
- $S_{0043231}(0043229) = w_{Is\_a} * S_{0043231}(0043231) = 0,8 * 1 = 0,8$
- $S_{0043231}(0043227) = w_{Is\_a} * S_{0043231}(0043231) = 0,8 * 1 = 0,8$
- $S_{0043231}(0005622) = w_{Part\_of} * S_{0043231}(0043229) = 0,6 * 0,8 = 0,48$
- $S_{0043231}(0005623) = w_{Part\_of} * S_{0043231}(0005622) = 0,6 * 0,48 = 0,288$
- $S_{0043231}(0043226) = \max(w_{Is\_a} * S_{0043231}(0043229) | w_{Is\_a} * S_{0043231}(0043227)) = \max(0,8*0,8|0,8*0,8) = 0,64$
- $S_{0043231}(0005575) = \max(w_{Is\_a} * S_{0043231}(0005623) | w_{Is\_a} * S_{0043231}(0043226))$

  $= \max(0,8 * 0,288 | 0,8 * 0,64) = 0,512$

On peut alors calculer la valeur sémantique du terme 0043231 notée SV(0043231) avec la formule suivante :

$$SV(a) = \sum_{t \in Ta} Sa(t)$$

d'où SV(0043231) = 7,112.

Sur la base de ces mesures Wang propose la formule suivante :

$$sim\ (a, b) = \frac{\sum_{t \in Ta \cap Tb}(Sa(t) + Sb(t))}{SV(A) + SV(B)}$$

Cette mesure présente l'avantage de tenir compte de l'information contenue dans un nœud donné (terme ou concept), de la position du nœud dans l'ontologie (sa position par rapport à la racine). Wang rajoute également un nouveau paradigme à savoir le poids de la relation parent-fils.

Pour la mesure de similarité entre gène voir l'annexe 7.

2. Production des réseaux gènes sur la base de la proximité sémantique

2.1. Notre approche basée sur la similarité sémantique

L'approche que nous proposons ici est une approche simple basée sur l'application d'une distance sémantique qui tient compte de la spécificité de GO. Dans GO comme cela est mentionné précédemment, un vocabulaire contrôlé a été développé spécialement pour décrire les processus biologiques (BP). Les produits de gènes, notamment de l'espèce humaine, ont été annotés par ce vocabulaire. Nous supposons ici que des gènes proches sémantiquement sur la base des annotations GO BP interviendraient dans les mêmes voies biologiques.

Notre démarche est basée sur deux étapes essentielles (Figure 41).

i. Le calcul de la similarité sémantique pour évaluer cette proximité entre gènes : Ici nous avons dans un premiers temps extrait la base d'annotations de GO. Puis nous avons filtrés les annotations de l'ontologie *Biological Process*. Nous avons par la suite implémenté un code Python pour le calcul de la similarité sémantique et la convertir en matrice de distance que nous utilisant pour l'étape de regroupement.

ii. Une étape de regroupement pour regrouper les gènes proches : Nous avons implémenté aussi un code Python qui permet de tester plusieurs algorithme de regroupement grâce au package Scikit[59].

→ Vers la fin du processus, nous voulons avoir des groupes de gènes qui peuvent reconstituer les voies biologiques en y incorporant les données d'interactions (répertoriées dans les bases de données).

2.2. Validation

Pour la validation de cette approche, nous avons proposé la démarche de la Figure 42. A partir de la base de données KEGG pathway[60], nous avons extrait neuf voies biologiques (voir Table 5) : des voies de signalisation cellulaires, des voies métaboliques et des voies de maintien. Nous avons décidé de tester deux distances sémantiques, GS2 (Ruths et al. 2009) et Wang (Wang et al. 2007) sur les annotations BP de ces gènes et de tester deux méthodes de regroupement. Les distances choisies ont été développées spécialement pour la mesure de la proximité sémantique sur la base de *Gene Ontology*.

[59] http://scikit-learn.org/stable/modules/clustering.html

[60] http://www.genome.jp/kegg/pathway.html

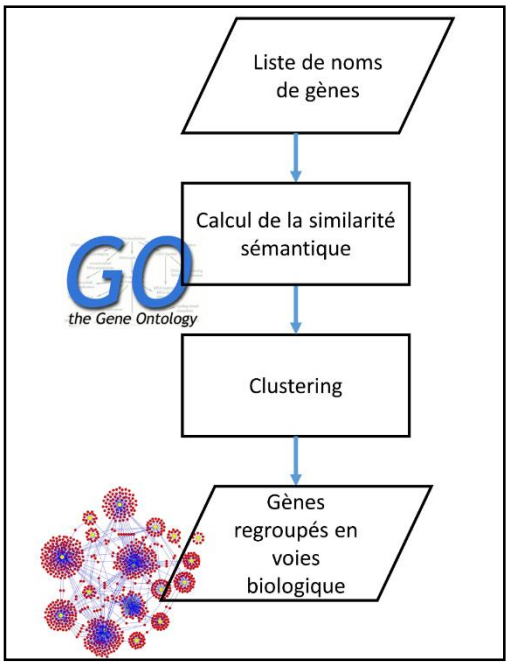


Figure 41 : Approche pour la prédiction de voies biologiques. Les cadres représentent les codes Python que nous avons développé les losanges représentent les entrées et sorties. A partir d'une liste de noms de gènes tirée au hasard, notre code génère les annotations GO dans l'ontologie *biological process*. Sur la base de ces annotations, les similarités sémantiques sont calculées et une matrice de distances est générée. Après une étape de regroupement *(clustering)* des gènes similaires.

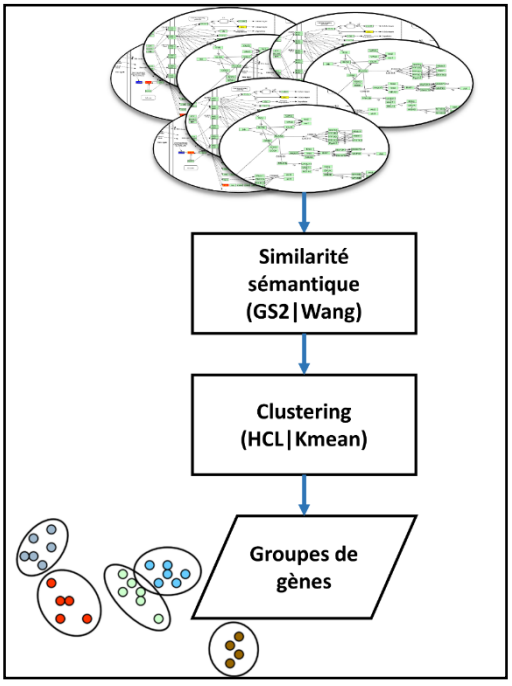


Figure 42. Démarche de validation de l'approche de modélisation des voies biologiques basées sur la similarité sémantique. Nous sommes partie d'un ensemble de voies biologiques répertoriées dans la base de données KEGG Pathways. Nous avons utilisé les codes implémentés sous Python pour le calcul de la similarité sémantique et l'implémentation des algorithmes de clustering. Nous avons comparés au final les groupes générées par rapport à nos données de départ.

Nous avons fait varier le nombre de clusters à la sortie et comparer les résultats à nos données de départs. Le nombre de clusters à la sortie varie entre 3 et 9 clusters pour chaque distance et chaque algorithme de regroupement, ce qui nous permet d'évaluer la sensibilité de la technique. Le but étant de voir si on arrive à dissocier les catégories ou bien même les voies biologiques avec une des techniques.

Table 5. Voies biologiques choisies pour la validation de la démarche de reconstitution des voies biologiques. Entre parenthèse, nous avons inscrit les diminutifs des voies biologiques que nous avons utilisés pour la suite.

Catégories	Voies biologiques KEGG	Nombre de produits de gènes
Métabolisme	Métabolisme des Acide Gras (Fatt)	6
	Voie du Pentose Phosphate (Pent)	18
	Métabolisme du Rétinol (Retinol)	19
	Métabolisme de la vitamine B6 (Vit B)	5
Système excréteur	Réabsorption de l'eau par régulée par la Vasopressine (Vaso Ab)	17
Cycle cellulaire	Voie de signalisation des estrogènes (Estro)	34
	Voie de signalisation des récepteurs des cellules T (T Cell)	63
	Voie de signalisation des chimiokines (Chemo)	49
Cycle cellulaire	Apoptose (Apop)	49

Dans la Figure 43 et la Figure 44 sont présentés les résultats observés pour la distance GS2 et correspondant à l'obtention de six clusters avec deux algorithmes de regroupement. Nous montrons ici les résultats de deux expériences pour une seule distance et six clusters à la sortie pour montrer le processus d'analyse, sachant que les résultats des autres expériences sont similaires.

Nous remarquons ici que les gènes appartenant à des réseaux différents sont distribués sur plusieurs clusters. Plusieurs raisons peuvent expliquer cette distribution hétérogène :

- Un gène peut être impliqué dans un ou plusieurs réseaux biologiques
- Les annotations utilisées pour les gènes sont très hétérogènes (niveau de précision variable)

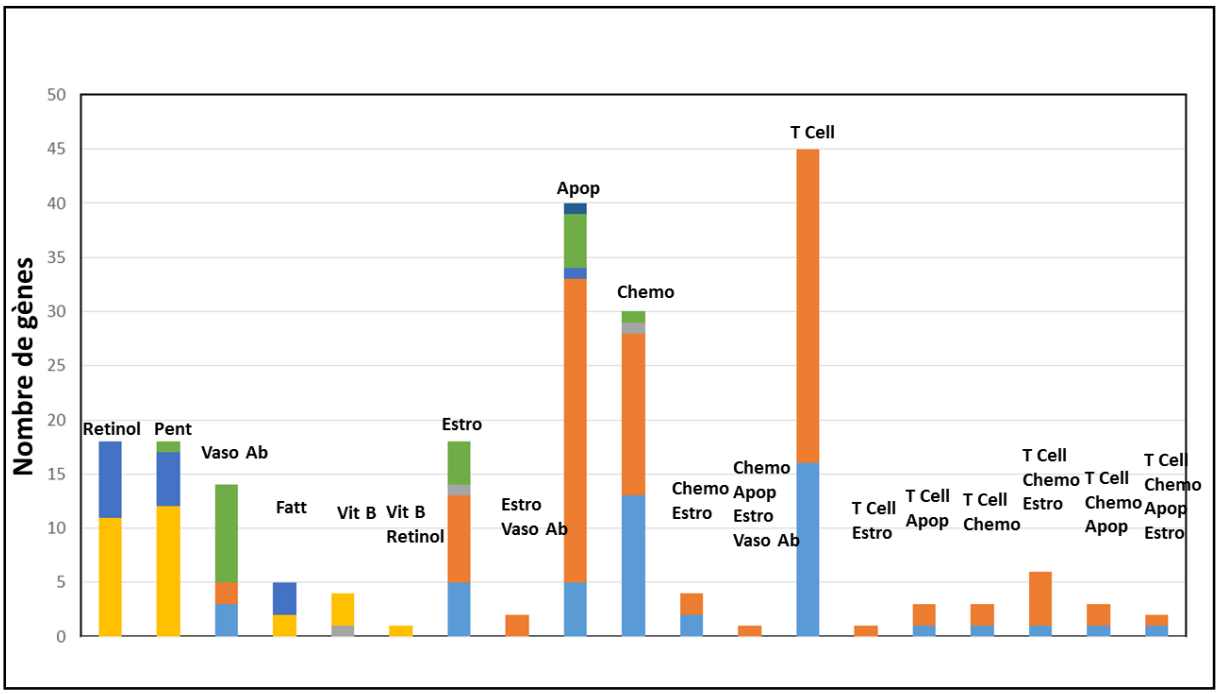


Figure 43. Résultat du regroupement pour la distance GS2 couplé à l'algorithme Clustering Hiérarchique. En haut, la représentation du résultat du regroupement obtenu par clustring hiérarchique, la barre horizontale rouge représente le site de coupure pour avoir 6 clusters et les différents clusters de gènes obtenus sont représentés par différentes couleurs. En bas, un histogramme représentant la distribution des gènes appartenant aux différents clusters sur les voies biologiques de la Table 5.

Pour remédier à ces deux inconvénients, nous avons adopté une nouvelle stratégie basée sur l'utilisation de la distance entres termes au lieu de la distance entre gènes pour l'évaluation de la proximité sémantique. Le but de cette stratégie est de voir si on peut identifier des clusters de termes GO spécifiques d'une voie KEGG. Nous avons décidé de prendre les voies KEGG une par une de procéder à une annotation des gènes qui la constituent, d'homogénéiser les annotations en ne prenant que les annotations appartenant au même niveau et de produire des clusters de termes GO et de voir s'il y a au moins un cluster qui regroupe la totalité des annotations ou au moins qui « absorbe » une grande partie du signal. On appelle ici signal, une annotation qui appartient à un cluster. Pour la notion de niveaux, nous avons adopté l'algorithme de Dijkstra (Dijkstra 1971) pour le calcul du plus court chemin. Nous considérons le niveau de précision comme le plus court chemin entre la racine et le terme en question. Cette méthode nous a permis d'identifier 11 niveaux : le niveau 0 est le niveau de la racine et le niveau 11 correspond au niveau de la feuille la plus basse.

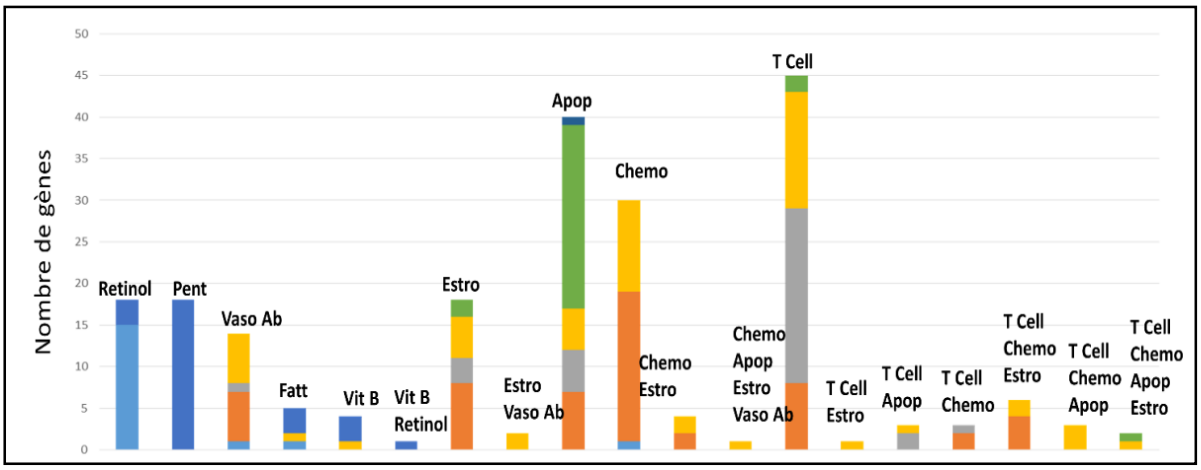


Figure 44. Résultat de regroupement pour la distance GS2 couplée à l'algorithme K-mean. **Dans cette expérience, nous avons généré six clusters. . Pour chaque cluster nous avons attribué une couleur. Chaque barre de l'histogramme représente le nombre de gènes appartenant à une ou plusieurs voies KEGG selon l'indication en haut de la barre. Sur cet histogramme nous avons représenté la distribution des gènes appartenant aux différents clusters générés sur les voies KEGG.**

Nous avons donc généré les annotations BP de notre groupement de gènes de départ, et nous avons fait une extension aux ancêtres. Nous considérons que si un terme annote un gène tous ses parents l'annotent aussi. Nous avons fait varier le niveau de précision de 5 à 7 car une grande partie des termes se trouvent entre ces deux niveaux. Notre hypothèse de travail est que les termes hauts dans GO (près de la racine) sont très généraux et pas assez informatifs ; nous aurons donc un échantillon assez uniforme ce qui va biaiser notre regroupement. Les termes très bas dans GO (prés des feuilles) sont très précis et donc annotent peu de gènes qui de plus seront très dispersés. C'est pourquoi nous avons opté pour des niveaux équidistants de la racine et des feuilles (entre 5 et 7). Nous avons implémenté cette approche sous Pythons. Nous présentant dans la Figure 45 les résultats de l'approche pour la voies de signalisions des récepteurs des cellules T. Dans cette expérience, nous sommes partis des 63 gènes de la voie de signalisation ; 24 gènes seulement possèdent au moins une annotation des niveaux 5 ou inférieurs dans l'ontologie *Biological Process* dans GO. L'expérience montre que les signaux de cette voie biologique et des gènes qui la composent sont répartis sur l'ensemble des clusters et ceci quel que soit la distance, la méthode de clustering et le nombre de clusters générés. Nous ne pouvons donc pas sur la base de cette approche regrouper les gènes qui appartiennent aux mêmes voies biologiques. Il en a été de même pour tous les autres réseaux biologiques, même en faisant varier la distance. Deux raisons peuvent expliquer ces résultats. Une première explication est la nécessité de filtrer les données ; en effet, le bruit de fond pourrait être responsable de l'étalement du signal. Un filtre en fonction de l'enrichissement et de la

pertinence de l'annotation pourrait améliorer le résultat en éliminant les annotations non pertinentes.

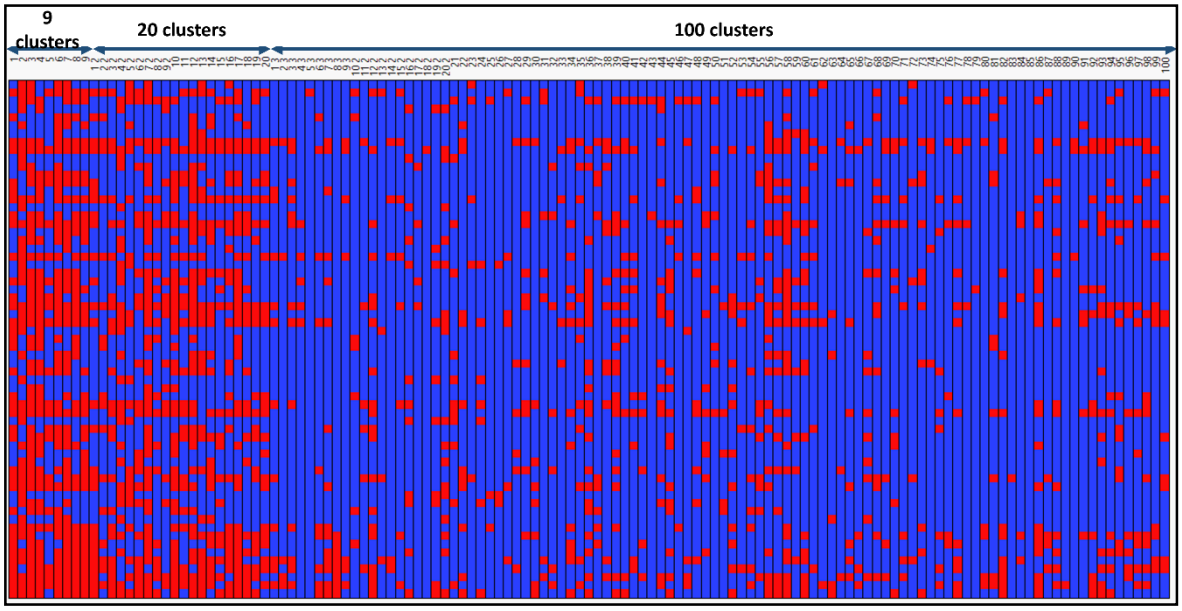


Figure 45. Distribution des signaux des gènes de la Voie de signalisation des récepteurs des cellules T dans les différents clusters. Chaque colonne représente un cluster et chaque ligne un gène. La coloration rouge représente au moins une annotation du gène sur la ligne qui appartient au cluster (en colonne). Pour cette expérience, nous avons fait varier le nombre de clusters générés (9, 20 et 100). Les flèches bleues indiquent les limites de chaque expérience. Pour cette figure, nous avons utilisé la distance GS2 en prenant en compte les termes GO de niveau 5 seulement.

Une autre explication serait l'utilisation la proximité sémantique basée sur les annotations dans l'ontologie « *Biological Process* » de GO comme source d'informations. En effet, une voie biologique est un ensemble de gènes en interaction pour assurer une fonction biologique bien déterminée. Si GO nous renseigne sur la fonction biologique, aucune information n'est donnée sur l'interaction d'où la nécessité d'introduire une autre source d'information sur l'interaction.

Nous avons donc pensé à un système de reconstitution des voies biologiques un peu plus complexe qui intègre deux notions : la proximité sémantique et l'interaction biologique. Pour cela nous avons entamé une collaboration avec le Laboratoire en Informatique en Programmation Algorithmique et Heuristique (LIPAH) de la Faculté des Sciences de Tunis. Que nous allons présenter dans ce qui suit/

C. Recherche de communautés pour la modélisation des voies biologiques

L'étude de l'interactome a permis de quantifier le nombre d'interactions binaires qui interviennent dans une cellule humaine à un instant « t » à plus de 130 000 (Venkatesan et al. 2009) et ce nombre se réfère uniquement aux interactions protéines-protéines (Figure 46).

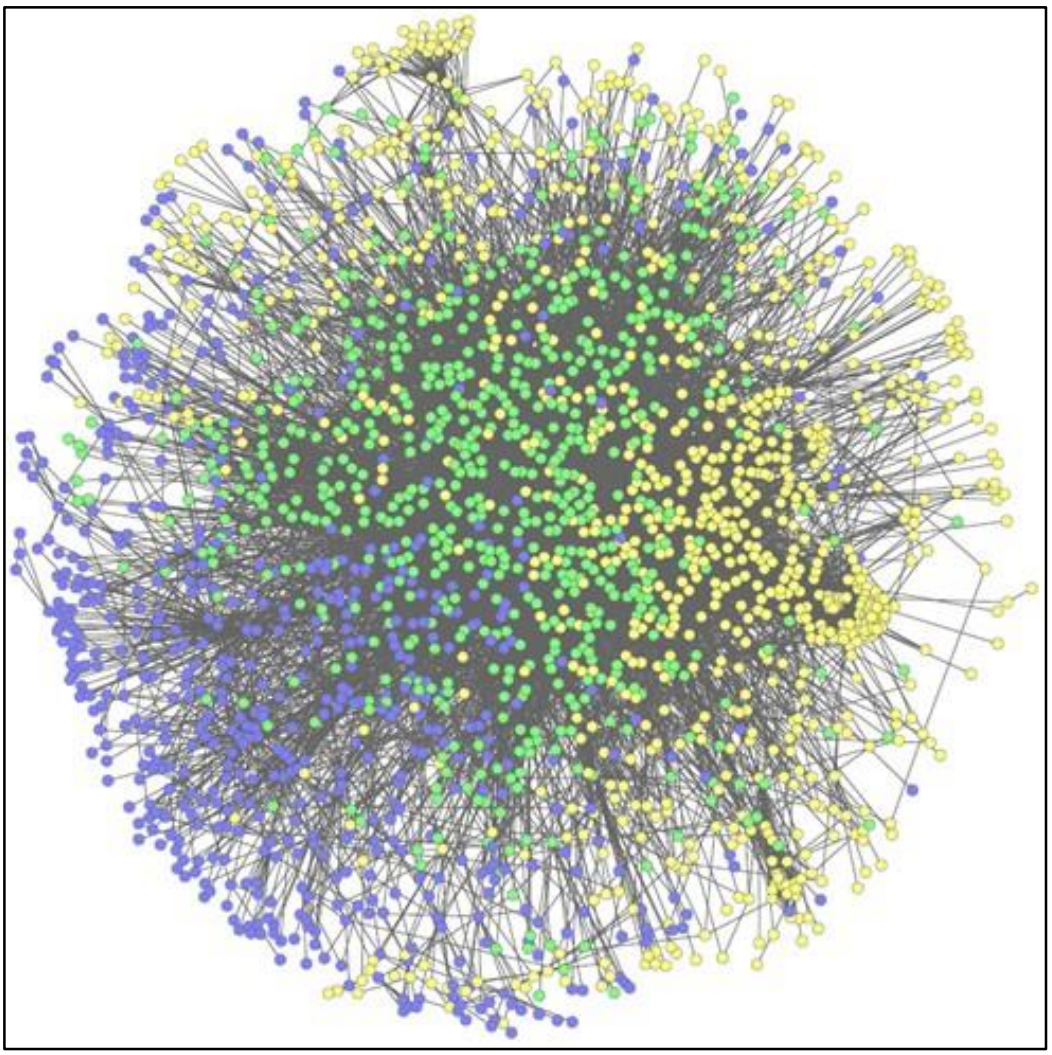


Figure 46. Modélisation de l'interactome humain par l'outil Cytoscape. **Dans cette figure, seules les interactions protéine-protéine sont modélisées. Les points représentent les protéines et les arrêtes représentent les interactions. D'après une présentation de ZHU FENG. Source : http://slideplayer.com/slide/7986778/.**

De ce point de vue, si on rajoute les autres interactions génétiques (ADN-ARN, ADN-Protéine, ARN-Protéine…), ça augmenterait la complexité du graphe. Ce graphe est composé de sous-ensemble d'éléments (nœuds) qui interagissent ensemble simultanément ou d'une manière séquentielle pour assurer une fonction biologique donnée. Ces sous-ensembles agissent

comme des communautés. Pour la reconstitution des réseaux biologiques, nous nous sommes positionnés dans une perspective de recherche de communautés dans un grand graphe complexe (l'interactome). Ici on appelle interactome l'ensemble des interactions biologiques impliquant l'ADN, l'ARN et les protéines dans une cellule. Pour la détection de communautés, nous avons combiné deux informations :

- La proximité sémantique : par calcul de la similarité sémantique en utilisant les GOA BP.
- Les données d'interactions : nous avons utilisé les données répertoriées dans la base de données db-string (Szklarczyk et al. 2015). Db-string[61] est une base d'interactions qui répertorie plus de 184 millions d'interactions produites par expérimentation et/ou criblage des bases de données et de la littérature pour plus de 2 000 organismes. Pour chaque interaction, on associe dans la base db-string un score de significativité basé sur la pertinence de la source.

1. Détection de communautés

Si la théorie des graphes a émergé depuis le 16ème siècle ce n'est qu'en 1927, que Stewart Rice a commencé les travaux sur la détection de communautés dans les graphes en faisant un regroupement manuel pour investiguer des blocks de votants (Sammut and Webb 2011). Dans les années 1930, la discipline d'analyse des structures des communautés commence à émerger. En 1955, la première étude à grande échelle pour l'analyse de communautés a été menée par Weiss et Jacobson pour la recherche de groupes de travail au sein d'une agence gouvernementale et sur la base de leurs travaux plusieurs approches ont été développées (Susskind et al. 2005). Navarro and Cazabet ont défini une communauté comme « *une structures mésoscopiques porteuses de sens* ». L'appellation détection de communautés a été inspirées de l'application que l'ont fait dans les grands graphes des réseaux sociaux néanmoins les champs d'applications sont plus étendu. De ce fait une communauté est un sous-ensemble de sommets qui peut être par exemple « *des concepts dans un graphe lexical, ou des thématiques dans un graphe de documents* » (Navarro and Cazabet 2010).

Dans la nature, les graphes n'ont pas une structure aléatoire. Les approches classiques pour la détection des communautés se basent essentiellement sur la topologie du graphe. Plusieurs approches ont été développées pour l'étude des structures (détection de communautés). La

[61] http://string-db.org/cgi/input.pl?UserId=aCHK9GItgY4F&sessionId=jzhNQPaBjN0M&input_page_show_search=off

détection de communautés dans les grands graphes est un problème NP-Complet. En informatique la complexité est une fonction théorique qui permet de modéliser la quantité de ressources nécessaire (temps et mémoire) pour exécuter un algorithme en fonction de la taille (n) des données de départ. On parle de problèmes NP si le temps d'exécution d'un algorithme pour aboutir à une solution est une fonction polynomiale. Un problème est noté NP-complet si le temps d'exécution pour aboutir à une solution tend vers l'infini. Parmi les solutions proposées pour la détéction de communautés, figurent les heuristiques dont les algorithmes génétiques (qui sont des méta-heuristiques qui se prêtent très bien à cet exercice) et la programmation dynamique (Guo et al. 2017). En termes de performances (temps d'exécution, consommation de mémoire) les heuristiques restent de loin les algorithmes les plus adéquats pour des graphes de tailles variables. En effet, dans notre cas, le nombre de gènes diffère d'un contexte à un autre et le nombre d'interactions est généré dynamiquement pour chaque contexte.

2. Algorithmes génétiques

2.1. Définition

Ce sont des méta-heuristiques appartenant à la classe des algorithmes évolutionnistes. Leur but est d'obtenir des solutions approchées lorsque la solution est inconnue ou lorsque la méthode pour l'obtention de la solution exacte est trop longue. Ce sont des algorithmes bio-inspirés et utilisant un champ lexical proche du jargon utilisé dans le domaine de la génétique (Holland 1992).

Les algorithmes génétiques ont été introduits par John Holland et son équipe de l'Université du Michigan en 1960. Ces recherches ont été publiées en 1975 dans le livre Adaptation in Natural and Artificial System. Néanmoins, on doit la popularité de cette approche à David Goldberg à travers son livre *Genetic Algorithms in Search, Optimization, and Machine Learning* en 1989 (Goldberg 1989).

2.2. Principe des algorithmes génétiques

Avant de décrire le principe de ces algorithmes, nous présentons tout d'abord la terminologie utilisée :

- Individu : les individus sont les solutions du problème d'optimisation. Ces solutions sont le plus souvent représentées sous la forme de vecteurs pour les nécessités du traitement. On appelle ce vecteur de solutions un chromosome.

- Population : c'est l'ensemble des chromosomes d'une même génération, qui sont habituellement d'une taille constante tout au long du traitement.
- Fitness : c'est un score pour évaluer la qualité d'un individu. Les algorithmes génétiques étant une technique d'optimisation, on recherche, selon l'objectif de l'étude, la qualité maximale ou minimale de la fonction à optimiser.
- Sélection : selon la qualité des individus, chacun se voit attribuer une probabilité (pourcentage de chances) d'être choisi. Cette probabilité correspond à l'importance relative de la qualité de l'individu par rapport à la qualité totale de la population.
- Reproduction : c'est une étape intermédiaire de l'algorithme génétique. C'est le croisement entre deux individus de la génération n-1 pour donner deux nouveaux individus à placer dans la nouvelle population de la génération n. Deux opérateurs sont responsables de l'étape :

  (i) L'recombinaison : durant cette étape chaque chromosome enfant reçoit la moitié des gènes des chromosomes parents. On applique ici un opérateur appelé « recombinaison » (*cross-over*).

  (ii) La mutation dont la probabilité est habituellement entre 0,5% et 5,0% du nombre d'individus. L'idée est de voir s'il y a des individus qui n'appartiennent pas aux chromosomes (à la solution proposée) et qui améliorent les performances (voir Figure 47 étape 3).

Le principe des algorithmes génétiques peut être résumé de la façon suivante (Figure 47) : Nous partons d'une population d'individus de base à n bits ou n caractères. Chaque élément représente un chromosome. (1) On va commencer par évaluer la qualité de chaque chromosome par le calcul du score de *fitness* (voir plus haut). (2) Pour la sélection, on fait un tirage au hasard de n/2 couples. (3) Chaque couple donne deux chromosomes fils par recombinaison (croisement). Ces étapes sont itérées. Le développeur peut définir un certain nombre d'itérations ou bien une condition d'arrêt.

2.3. Algorithmes génétiques et détection de communautés

Plusieurs auteurs ont utilisé cette approche pour l'étude de la structure des communautés dans les graphes complexes. Grâce à leurs travaux, Talbi et Bessière (Talbi and Bessière 1991) ont démontré que le problème de partitionnement de graphe qui est un problème NP-complet pouvait être résolu efficacement grâce aux algorithmes génétiques.

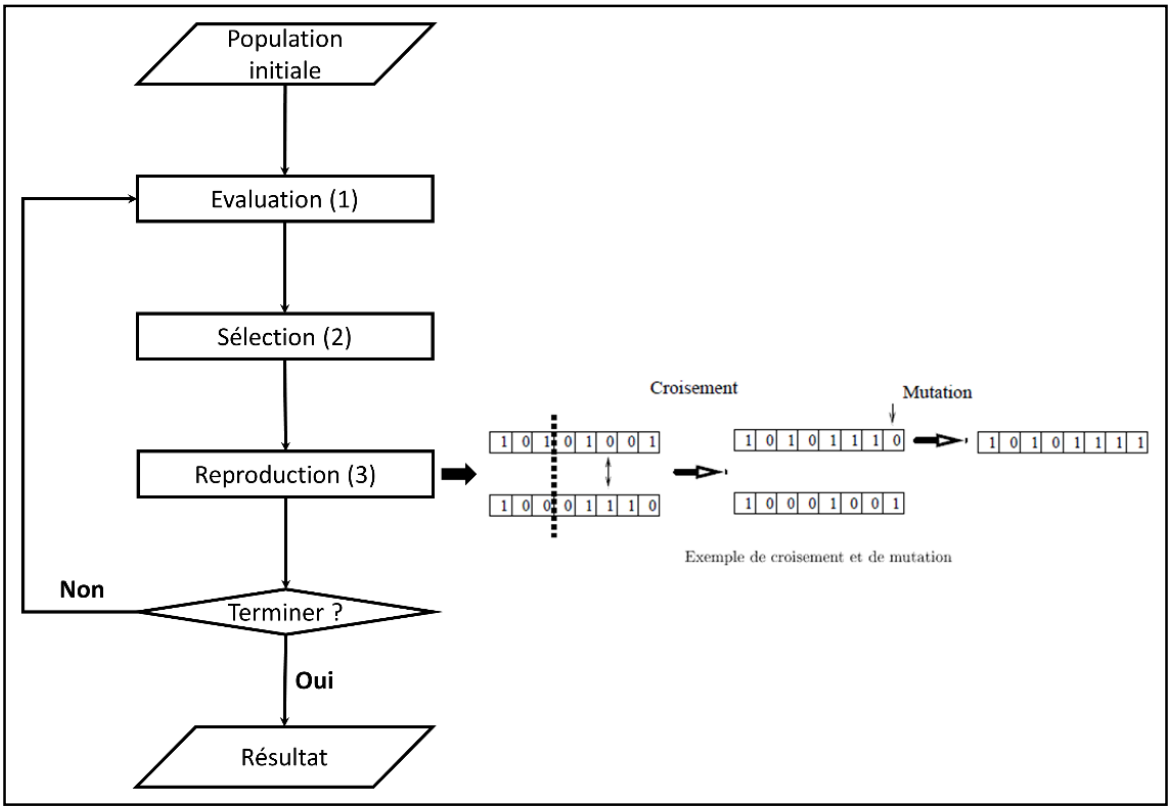


Figure 47. Principe de fonctionnement d'un algorithme génétique. **L'algorithme se déroule en quatre étapes de manière itératives, jusqu'à atteindre une condition d'arrêt. A droite, nous avons modélisé les deux opérateurs (recombinaison et mutation) de l'étape de reproduction. Les gènes ici sont modélisés par les 0 et des 1 et ils sont contenus dans des chromosomes.**

Tasgin (Tasgin et al. 2007) a produit une méthode où il prend comme fonction fitness la modularité[62]. La méthode a été validée sur des données sociologiques. En 2008, Pizzuti (Pizzuti 2008) a présenté GA-Net un algorithme codé sous MATLAB[63] pour la détection de communautés dans les réseaux sociaux. Cet algorithme utilise une représentation basée sur les matrices d'adjacence pour développer une nouvelle mesure de *fitness* appelée « score de communauté ». L'avantage de cette méthode est la réduction du nombre de calculs « inefficaces ». Plus récemment, des approches hybrides ont été, comme celle présentée par Li et Liu (Li and Liu 2016). Cette approche combine les algorithmes génétiques et les modèles multi-agent, ce qui permet des performances de détection de communautés plus sensibles tout en améliorant la vitesse de détection et la stabilité de l'algorithme.

[62] C'est une mesure pour quantifier l'aptitude d'un graphe à être divisé en modules. Les réseaux avec une haute modularité sont des graphes avec une grande connectivité entre les nœuds du même module et une faible connectivité avec les nœuds des autres modules au sein du même graphe.

[63] https://fr.mathworks.com/products/matlab.html

3. Notre approche pour la modélisation des voies biologiques

Au début du chapitre nous avons défini une voie biologique comme étant un ensemble de gènes en interactions et proches sémantiquement qui assurent une fonction biologique. La revue de la littérature montre qu'à un instant t une cellule humaine est capable de faire plus de 130 000 interactions protéiques, si on rajoute à cela les interactions gène-protéine, ARN-protéine, ceci constitue un grand graphe complexe d'interactions génétiques. Ce qui nous amène à penser qu'une voie biologique constituerait une communauté au sein de ce graphe. Notre but ici est de regrouper les gènes susceptibles d'appartenir aux mêmes voies biologiques (communauté). L'approche, co-développée avec l'équipe LIPAH, a été présentée dans le rapport de master de Mlle Marwa Ben M'Barek (BEN M'Barek 2016). Elle propose un algorithme génétique qui utilise une fonction *fitness* combinant la proximité génétique GS2 (Ruths et al. 2009, p. 2) et le score d'interaction de la base de données db-String (Szklarczyk et al. 2015). Nos travaux ont consisté à la validation de cette approche. Nous voulions ici proposer une approche généraliste qui part d'une signature de gènes quelconque (la signature peut être produite à partir d'expérience de transcriptomique, ou bien de contextualisation…) et de regrouper les gènes en fonctions des voies biologiques auxquelles ils seraient susceptibles d'appartenir. Le but étant de pouvoir intégrer cette étape après contextualisation. En effet la contextualisation permet d'apporter un complément d'informations sur les conditions macroscopiques d'expression (l'état pathologiques et le compartiment anatomique) et les conditions microscopiques (population cellulaires). Reste maintenant comprendre les mécanismes biologiques mis en jeux. La première étape consiste à regrouper la liste de gènes en voies biologiques. Pour mener cette tâche nous nous sommes proposé de développer un package sous Python. Le processus de validation est en cours c'est pour cette raison qu'on n'aborde pas l'exemple du vieillissement du système immunitaire ici.

3.1. Le principe de l'approche de modélisation des gènes appartenant aux mêmes voies biologiques

On part d'un ensemble de gènes (une signature de gènes contextualisés, une signature de gènes issue d'une expérience des transcritomoque) pour les classer en voies biologiques. (1) Pour chaque gène ainsi identifié, la première étape consiste à récupérer la liste de ses interactions possibles (tous les gènes avec lesquels ce gène interagit). Cette liste contient les scores d'interactions (ou scores de significativité calculés sur la base de la source de l'information) et le type d'interaction (protéine-protéine ou bien une réaction de régulation moléculaire). Notre

liste de départ peut donc être enrichie à partir de db-string. (2) L'étape suivante consiste à calculer les similarités sémantiques GS2 gène à gène. Une étude comparative entre des méthodes de calcul de distance (distance de Lin, distance de Wu Palmer, la distance Wang, GS2) a été au préalable effectuée avant de choisir cette dernière. (3) A partir de ces données, un score de *fitness* entre gènes est calculé d'abord deux à deux et (4) des listes de gènes sont générées aléatoirement qu'on appelle « *biological pathways* » ; pour chaque *pathway*, un score de *fitness* global est calculé et, sur la base de ce score, les *pathways* sont triés. Le nombre de gènes par *pathway* est fixé par l'utilisateur comme paramètre de l'algorithme. Dans ce qui suit nous avons considéré que le nombre arbitraire de 10 gènes est représentatif d'une voie biologique et surtout qu'il est gérable pour une étape de validation par un expert. (5) L'étape de reproduction définie plut haut, prend en compte le paramètre de recombinaison qui est de 50% des gènes de chaque *pathway* et le paramètre mutation qui est de deux gènes avec les scores de fitness les plus faibles qui sont éliminés pour les remplacer par d'autres gènes de db-string et n'appartenant pas à la signature de départ. Pour définir ces deux paramétres nous avons suivi les recommandations de (Tasgin et al. 2007). Cette étape est itérée vingt fois car nous avons observé qu'après vingt itérations, le résultat final se stabilise (Figure 48).

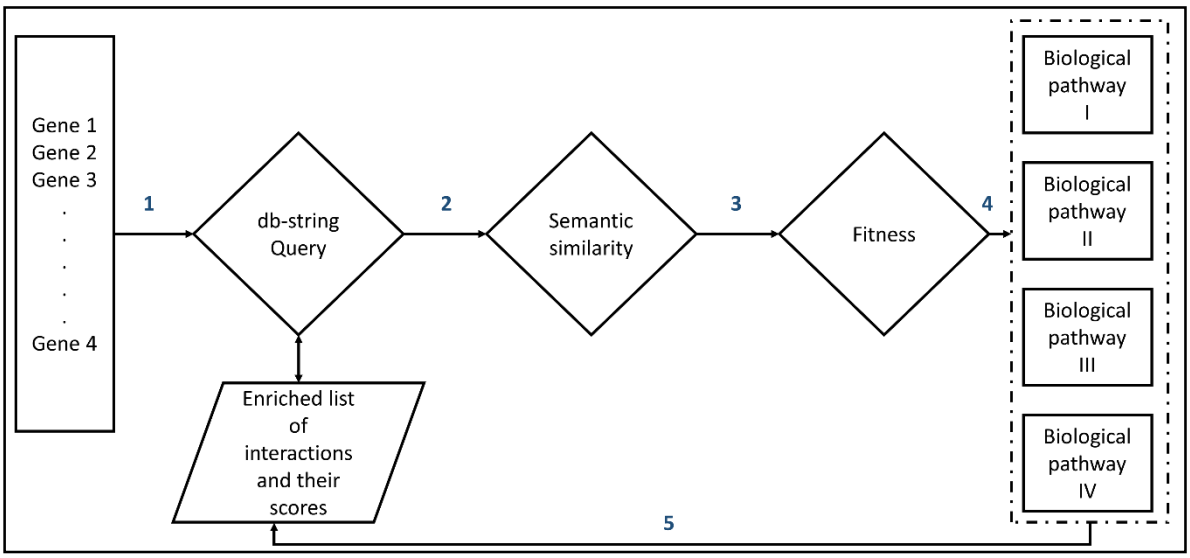


Figure 48. Schéma récapitulatif de l'approche utilisée pour la modélisation des voies biologiques. 1) A partir d'une liste de gènes (signature de gêne), nous lançons une requête dans la base de données db-string. Pour chaque gène de notre liste de départ, nous récupérons tous les gènes avec lequel il interagit et un score d'interaction (calculé par db-string). 2) Nous calculons la similarité sémantique GS2 entre les différents gènes de liste de départ. 3) Nous combinons la similarité sémantique et le score db-string pour le calcul de la fonction fitness. 4) Les gènes sont regroupés en voies biologiques (*biological pathway*) potentielles sur la base de cette fonction. 5) C'est l'étape de mutation ou l'on enrichit notre liste de gènes de départ par des gènes issus de la requête db-string.

3.2. Notre approche pour la validation de l'algorithme génétique de modélisation des voies biologiques

Pour la validation de cet outil nous avons proposé l'approche présentée dans la Figure 49. Les gènes choisis pour la validation ont été pris au hasard à partir de la base de données KEGG.

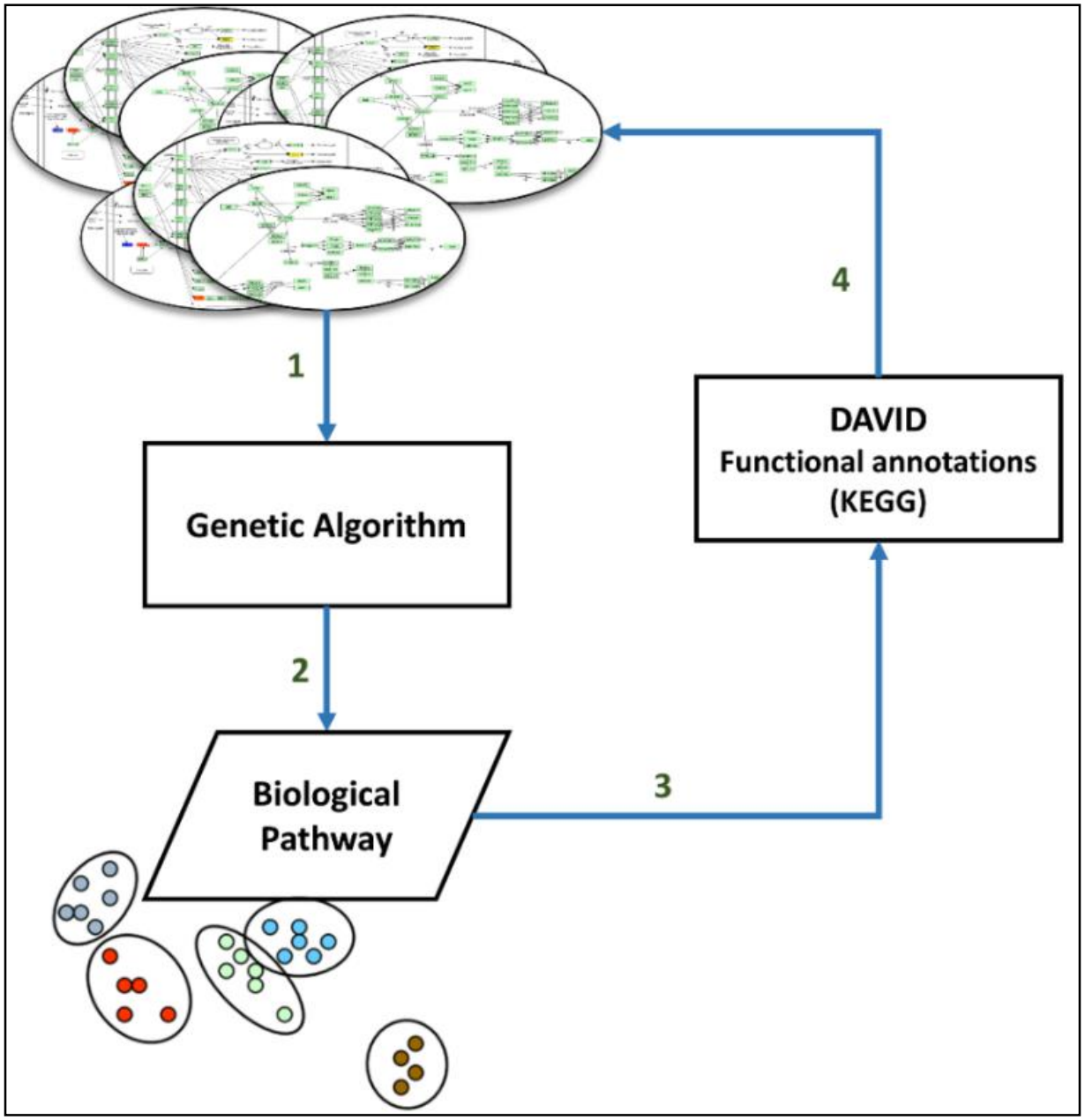


Figure 49. Schéma expérimental pour la validation de l'outil de reconstitution des voies biologiques. (1) Nous partons d'une liste de gènes qui appartiennent à des voies biologiques connues à partir de la base de données KEGG (Table 5). (2) Nous appliquons notre approche sur cette liste de gènes (indépendamment de leur classement initial) pour valider le classement en groupe « Biological pathways ». (3) Le résultat obtenu est annoté par l'outil DAVID (Dennis Jr et al. 2003). Nous avons utilisé les API de DAVID pour automatiser le processus et avons paramétré la requête pour n'avoir que les annotations KEGG. (4) Pour l'évaluation de l'outil, nous comparons les résultats obtenus aux réseaux KEGG de départ.

Table 6. Table récapitulative des 30 gènes utilisés pour l'expérience de validation appartenant initialement à 3 voies biologiques différentes

Nom du réseau	Gènes
Ascorbate And Aldarate Metabolism	*'UGT1A10','UGT1A8','UGT1A7','UGT1A6','ALDH1B1','UGT2B28',* *'ALDH2','UGT1A5', 'MIOX','UGDH'*
T Cell Receptor Signaling Pathway	*'MALT1','PAK4','AKT3','IL4','IL5','FOS','CBL','SOS1','MAPK1','PTPN6'*
Oocyte Meiosis	*'CDC16','CALM1','RPS6KA6','SLK','IGF1R','ANAPC5','AURKA','CCNE2','SMC3','CCNB2'*

3.3. Résultats de la validation

Notre objectif pour la validation est de voir si on arrive à reconstruire des voies biologiques connues. Pour cela nous avons choisi de partir de voies biologiques répertoriées dans la base de données KEGG Pathway. Nous avons choisi : une voie métabolique *Ascorbate And Aldarate Metabolism*, une voie du cycle cellulaire : *Oocyte Meiosi*, et une voie immunitaire : *T Cell Receptor Signaling Pathway*. Le choix des voies biologiques s'est fait suite à l'analyse des résultats d'annotation des signatures de gènes contextualisées. Notre algorithme doit être paramétré en prenant en entrés le nombre de voies biologiques et le nombre de gènes par voies. Nous avons pris 30 gènes au hasard 10 gènes de chaque voie biologiques KEGG (Table 6). Nous avons paramétré notre algorithme pour avoir 10 voies biologiques probable se composant chacune de 10 gènes. Nous avons récupéré le résultat. Nous avons remarqué que 5 voies probables sont répétées, nous avons donc éliminé les voies redondantes. Nous avons annoté les gènes des 5 voies restantes en utilisant l'outil DAVID disponible en ligne contre la base de données KEGG pathway pour voir si on arrive à retrouver les voies de départ. Le résultat est présenté dans la Table 7. D'abord sur les 5 voies biologiques générées deux voies sont annotées *Ascorbate And Aldarate Metabolism*, deux voies sont annotées *Oocyte Meiosis* et une voie est annotée *T Cell Receptor Signaling Pathway*, donc les voies biologiques probables correspondent aux voies biologiques KEGG de départ. Ensuite pour les gènes des voies biologiques probables le taux de gènes appartenant aux voies biologiques effectives varient entre 40% et 70% avec un taux moyen de 47%. Pour conclure à cet étape de développent on peut dire que par rapport aux méthodes proposés dans le paragraphe « B » de ce chapitre il n'y pas de perte du signal car on arrive à retrouver nos voies biologique KEGG de départ, reste a améliorer le taux d'erreur.

Table 7. Table récapitulative des voies biologiques regroupées par notre algorithme génétique. Le résultat est trié par le nombre de gènes retrouvés dans les voies KEGG.

Voies biologiques modélisées	Score fitness	Nom du réseau KEGG	Nombre de gènes dans le réseau KEGG	%	P-value	Taux d'enrichissement
'UGDH', 'UGT1A5', 'ALDH2', 'AURKA', 'UGT1A8', 'UGT1A7', 'UGT2B28', 'ANAPC5', 'SLK', 'MIOX'	0,62	*Ascorbate And Aldarate Metabolism*	7	70	1,64E-13	179,15
CALM1', 'IGF1R', 'IL5', 'PAK4', 'PTPN6', 'SLK', 'RPS6KA6', 'ANAPC5', 'UGDH', 'SMC3'	0,65	*Oocyte Meiosis*	6	60	1,07E-07	38,04
ALDH2', 'UGT1A8', 'SOS1', 'SMC3', 'UGT1A7', 'UGT2B28', 'UGT1A10', 'MIOX', 'AKT3', 'CCNE2'	0,63	*Ascorbate And Aldarate Metabolism*	6	60	7,68E-11	153,56
'SLK', 'CALM1', 'UGT2B28', 'CBL', 'SOS1', 'IL5', 'IGF1R', 'MALT1', 'UGT2B28', 'UGT1A5'	0,76	*T Cell Receptor Signaling Pathway*	4	40	0,000171	29,82
'ALDH1B1', 'BUB1B', 'SYCE2', 'CDC16', 'CALM1', 'PAK4', 'SYCE2', 'ANAPC5', 'UGDH', 'SMC3'		*Oocyte Meiosis*	4	40	0,000128	31,70

D. Discussion, modélisation des voies biologiques

Le criblage de la littérature et l'analyse des signatures des gènes montre que les processus biologiques mis en jeux différent d'une signature à une autre (voir le chapitre précédent). Notre objectif derrière cette étape du projet est de pourvoir modéliser les voies biologiques d'une signature de gènes. Nous avons commencé par définir une voie biologique comme étant un ensemble de gènes proches sémantiquement et qui interagissent ensemble pour mener une fonction biologique. La première étape de cette modélisation est le regroupement des gènes susceptibles d'appartenir à la même voie. Plusieurs approches sont disponibles pour la modélisation des réseaux de gènes on peut citer :

- La modélisation statistique des données d'interactions combinées aux données issues des bases de données d'interactions (Segal, Wang, et Koller 2003)
- Le criblage de la littérature (Frisch et al. 2009)
- La détection de communautés basée sur la structure des graphes (Wilson et al. 2016)

De notre côté nous avons opté pour une approche basée sur l'information. Dans un premiers temps nous avons développé une approche basée sur la proximité sémantique (un paramètre unique couplé à un algorithme de classification par regroupement (*clustering*) cette approche n'a pas été informative limité par l'utilisation d'un seul paramètre. Pour cela nous avons fait le choix d'une approche plus riche en information et qui nous permet de considérer plus qu'un paramètre à la fois. Dans une collaboration avec le LIPAH, nous avons pu grâce un algorithme de classification (algorithme génétique) combiner deux paramètres : la proximité sémantique et les données d'interaction. Le principe de l'approche des algorithmes génétiques déjà publiée pour la détection des communautés au sein des graphes (Tasgin et al. 2007, Guerrero et al. 2017) est résumé dans la Figure 50. Le même principe a été utilisé pour la détection de groupes de gènes spécifiques à des pathologies (Muraro et Simmons 2016).

De notre côté nous avons opté pour une autre exploitation de l'approche des algorithmes génétiques en vue de la détection de voies biologiques. En effet les données d'expression de gènes et d'interactions comme ceux utilisées par certains auteurs (Muraro et Simmons 2016) sont complexes et peu exploitables pour une approche généraliste comme celle que nous avons présentée ici surtout vue la taille et le coût de ces données qui varient aussi en fonction du contexte d'expression. A cet effet nous avons exploité l'approche des algorithmes génétiques

en la combinant à la méthode de proximité sémantique entre gènes d'une part et aux informations relatives aux données d'interactions d'autre part. Dans l'analyse classique utilisant des algorithmes évolutionnistes, on se base sur la structure des graphes pour la recherche des communautés, alors seuls les nœuds fortement interconnectés seront considérés comme une communauté pouvant potentiellement constituer une voie biologique.

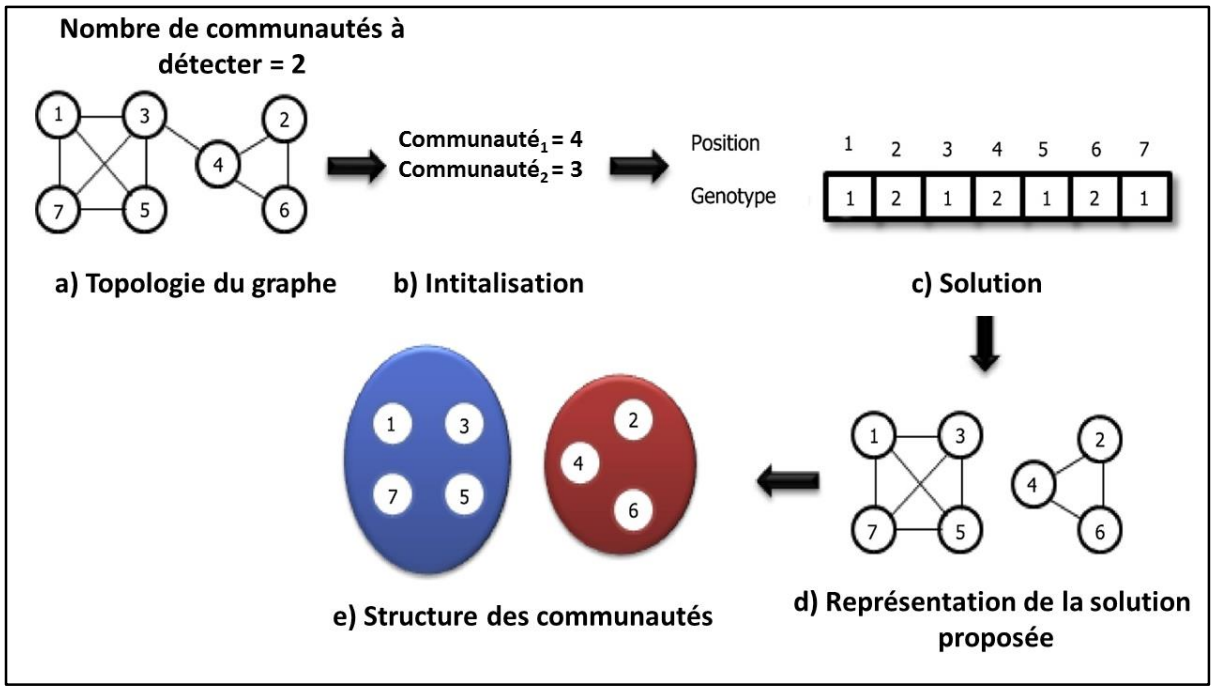


Figure 50. Principe de la détection de communauté dans les graphes sur la base de la structure (Guerrero et al. 2017). Source : http://www.sciencedirect.com/science/article/pii/S0925231217308664.

Dans notre analyse nous nous basons sur deux niveaux d'informations :

- d'une part la proximité sémantique calculée à partir des annotations *Biological Process* de *Gene Ontology* (Voir le paragraphe 3).
- D'une autre part sur la significativité de l'interaction en fonction de sa redondance dans la littérature et des bases de données.

L'exploitation innovante des algorithmes génétiques dans le présent travail a consisté à combiner ces deux paramètres pour modéliser des communautés que nous interprétons comme des voies biologiques ce qui permet de révéler de nouvelles interactions non décrites. Néanmoins on pourra dans une étape prochaine intégrer aussi les données de co-expressions pour améliorer les performances de notre approche. La collaboration avec le laboratoire LIPAH nous a permis de développer un outil sous Python utilisant un algorithme génétique pour cette

tâche. Dans notre approche de développement nous avons voulu concevoir un outil généraliste qui pourra être utilisée pour l'analyse des signatures de gènes quel que soit son origine (signature contextualisée, ou bien directement issue d'une expérience de transcriptomique). Le résultat de la validation montre que l'outil permet de retrouver toutes les voies KEGG de départ. Le taux des gènes bien classés varie entre 40% et 70%. Donc l'adoption de cette stratégie permet de classifier les gènes en voies biologiques probable bien que d'autres tests de validations sont en cours avec une masse de gènes en entrée plus importante, pour pouvoir entre autre pouvoir calculer la spécificité et le taux de rappel. Néanmoins contrairement à l'utilisation des distances sémantiques suivie d'un algorithme de regroupement qu'on a présenté dans le paragraphe B elle assure de meilleures performances. En effet nous avons vu que l'utilisation de la proximité sémantique couplée à un algorithme de classification ne permet pas de regrouper les gènes en voies biologiques. Entre la première expérience ou nous avons utilisé la proximité sémantique GS2 (Ruths et al. 2009) suivie d'un algorithme de regroupement soit clustering hiérarchique (Johnson 1967) soit le k-mean (Hartigan and Wong 1979), nous avons apporté une nouvelle information qui porte sur les interactions gène a gène. Cet apport d'information a permis d'améliorer les performances du regroupement même si on a utilisé un nouvel algorithme (l'algorithme génétique). Ceci conforte notre hypothèse de départ pour les voies biologiques selon laquelle une voie biologique est un ensemble de gènes proche sémantiquement et qui interagissent ensemble pour assurer une ou plusieurs fonctions biologiques (Kanehisa et al. 2017). Cette définition, des voies biologiques, vient en contradiction avec les approche rapporté par la littérature. En effet Guo et coll. proposent un modèle basée uniquement sur la proximité sémantique (Guo et al. 2006). Néanmoins ce modèle basée sur la proximité sémantique permet uniquement de prédire uniquement si deux protéines appartiennent ou non à la même voie biologique. Ceci ne permet pas répondre à notre besoin de départ qui de modéliser les voies biologiques. La première étape consiste à regrouper les gènes à partir d'une signature (transcriptomique, ou contextualisée) pour pouvoir par la suite contextualiser. Comme mentionné précédemment d'autres étapes de validations sont à prévoir pour cette méthode de regroupement notamment en pérennant en entrés des gènes appartenant à un nombre plus important de voies *KEGG* ceci permettra de calculer la spécificité et la sensibilité de la méthode en faisant varier la taille des voies biologiques à la sortie et le nombre de gènes regroupés à chaque fois. Pour l'étape de modélisation proprement dites il faut aussi intégrer un outil graphique qui prend en compte la nature de l'interaction (stimulation, répression de l'expression, ou bien une interaction protéique) ces informations sont disponible aussi sur la base de données db-string (Szklarczyk et al. 2015). D'un point de vue économique aussi

l'approche présentée ici permet de comprendre la nature des interactions et d'analyser les processus biologiques mis en jeux à travers les voies biologiques, en utilisant des ressources indexées et disponibles gratuitement (db-string et *Gene Ontology*). En effet vue le nombre de signatures de gènes contextualisées l'utilisation d'approches expérimentales telles que celle présentées par Roy et coll. et Anderson et coll. (Roy et al. 2013, Anderson et al. 2016), ou ils modélisent un interactome à partir de données transcriptomiques, ont un coût matériel et humain important et il est impensable, étant donné le nombre de signatures générées, de développer ces approches pour l'étude des listes de gènes contextualisés.

Donc pour conclure, nous avons présenté ici deux approches pour la modélisation des voies biologiques la première approche se base uniquement sur la proximité sémantiques et la seconde sur la proximité sémantique et les scores d'interactions. La deuxième approche permet de regrouper les gènes en fonction des voies biologiques probable. L'outil est développé sous Python. C'est un outil généraliste qui permet d'analyser n'importe quelle signature de gènes (trasncriptomique ou contextualisée) mais qui a été développé dans la perspective de compléter OntoContext pour l'analyse des processus biologiques mis en jeux dans chaque contexte biologique identifié.

Discussion Générale & Conclusion

La biologie expérimentale couvre un large spectre de disciplines, allant de la biologie cellulaire à l'immunologie et aux neurosciences (Weber 2014). Cette discipline scientifique est basée sur les observations et l'expérimentation et

> « *implique souvent des descriptions qualitatives (mais parfois aussi quantitatives) des mécanismes et des processus* » (Weber 2014)

Cette étape de description intervient durant la phase d'interprétation des résultats expérimentaux et c'est d'autant plus vrai dans le cas de l'interprétation des résultats des expériences sur le transcriptome. L'objectif de cette thèse est de présenter des outils d'aide à l'interprétation des données d'expression. Nous présentons ici une approche basée sur les méthodologies et les outils des sciences de l'information pour cette tâche. Si la notion d'information biologique en général a été adoptée depuis longtemps par la communauté scientifique en biologie (Spengler 2000), certains auteurs vont jusqu'à définir la bioinformatique comme la science de l'information biologique (Heidorn et al. 2007), ce concept d'information biologique est remis en cause par les spécialistes de la théorie de l'information (Longo et al. 2012). Une théorie qui dans son approche de l'information fait abstraction du sens (Floridi 2005, Leleu-Merviel and Useille 2008). Or, on a vu dans l'introduction que l'information est définie comme un ensemble de données qui ont un sens. Le sens, quant à lui, est défini en fonction du contexte. Selon Parrini-Alemanno (Parrini-Alemanno 2007), c'est le contexte

> « *qui fait que les individus ne sont pas considérés comme des variables mais comme un tout* ».

C'est pour cette raison que, dans toutes les autres approches de l'information, la notion du contexte a été liée au concept « information ». Parcontre, la théorie de l'information définit l'information comme un pattern ; elle s'intéresse ainsi à ce qui porte l'information et n'aborde pas l'aspect sémantique (Leleu-Merviel and Useille 2008).C'est une des limites de cette théorie. Longo et coll. vont jusqu'à dissocier la fonction biologique (le sens) du support (séquences nucléiques) (Longo et al. 2012). Pour l'information biologique en général, et l'information génétique en particulier, ce sont les séquences, particulièrement l'ADN, qui sont porteuses de l'information biologique. Cette information est exprimée en fonctions biologiques. Cette expression dépend d'un contexte d'expression (Attwood 2000). Justement, l'argument mis en avant par Longo et coll. est la stochasticité des mécanismes d'expression pour la remise en

cause de la notion d'information génétique. D'un autre côté selon la définition même de l'information, on ne peut pas la dissocier du contexte qui est à l'origine de la stochasticité de l'information biologique. On peut conclure que la séquence génétique (données), s'exprime en voies et processus biologiques (sens), sur la base du contexte d'expression.

La première problématique à laquelle il fallait répondre au court de ce projet a été de définir le contexte d'expression des gènes. La biologie de développement, adopte aussi cette approche de la relation gène-fonction (Orgogozo et al. 2015). Dans ce projet nous avons défini le contexte d'expression comme suit :

$$Contexte = population\ celluaire + compartiment\ anatomique + état\ pathologique$$

La deuxième problématique du projet a été de construire les outils technologiques qui permettent de définir ce contexte et d'identifier les voies biologiques mis en jeux. Une des contraintes de développement est l'automatisation du processus vu le développement technologique des techniques d'étude du transcriptome (Les puces à ARN et le séquençage à haut débit) comme le montre la revue de la bibliographie. Nous avons développé une approche à deux étapes. La première consiste à définir les contextes d'expression des gènes à partir de la littérature biologique. Cette question de contextualisation, comme nous l'avons définie, a été abordée par d'autres outils. Dans les ontologies biologiques par exemple, on peut trouver des liens entre plusieurs ontologies : par exemple, la *Cell Ontology* (Bard et al. 2005) contient des fragments de la *Gene Ontology* et de *UBERON Ontology* et inversement pour *l'UBERON Ontology* (Mungall et al. 2012). Ces liens peuvent être traduits par des relations entre les différents concepts des différentes ontologies. Ces insertions ont été imposées par l'initiative OBO (Smith et al. 2007) davantage pour des raisons d'interopérabilité que pour la contextualisation des gènes et de leurs produits. Le nombre de relations entre différents concepts des différentes ontologies est insuffisant - et reste très limité par rapport à la littérature – pour des études intégratives. Des projets sont en cours de réalisation sur la base de profils d'expression en fonction des populations cellulaires obtenues à partir de données expérimentales (Vempati et al. 2014) ou en fonction du tissu (Liu et al. 2008) en criblant les bases de données. Néanmoins, relier ces données reste difficile pour définir un contexte d'expression comme présenté dans ce projet pour des raisons d'interopérabilité aussi la question du coût économique de ces initiatives est une des grandes limites de ces initiatives. L'outil OntoContext développé dans ce projet permet quand lui des cribler la littérature pour l'identification de base de signatures de gènes contextualisée. En effet la littérature est une source d'informations disponible et surtout gratuite. MEDLINE est une base de données

bibliographique qui contient plus de vingt-cinq millions de références dont les résumées sont disponibles gratuitement, ce qui permet de réduire le coût économique. Le criblage de la littérature présente aussi un avantage quantitatif quant aux nombre de contextes qu'on peut générer, car la littérature constitue jusqu'aujourd'hui la source la plus importante de diffusion des études scientifiques. Dans ce projet le criblage d'un corpus de résumé d'article liée au vieillissement du système immunitaire a permis d'identifier plus de 2000 concepts cellulaires, 400 concepts anatomiques et de 3000 concepts pathologiques. D'autre outils de contextualisation ont été aussi rapporté dans la littérature tel que ANNI (Jelier et al. 2008) par exemple permet de lier les gènes par rapport à des concepts (soit pathologiques, soit cellulaires, soit anatomiques) sur la base de la littérature. Cet outil permet de lier les concepts deux à deux seulement et se base sur un corpus pré-annoté ce qui pose un problème pour la mise à jour de l'outil et sa personnalisation. En effet un des avantages de l'utilisation OntoContext, a été développé dans le but de cribler un corpus de texte grâce au module *annot*, la contextualisation se fait par entrecroisement des concepts identifiés par le module *crisscross*. On peut ainsi faire entrecroiser trois concepts des trois ontologies prédéfinis et avoir en résultat une liste de noms de gènes et de leurs produits contextualisée. L'utilisation des ontologies comme bases d'annotation a été appuyée par l'effort de standardisation qui a été entrepris par la communauté d'experts. L'outil NCBO Annotator (Jonquet et al. 2009) quant à lui permet d'annoter les ressources textuelle en fonction des ontologies répertoriées dans le portail BioPortal, nous démontrons ici qu'OntoContext a de meilleurs performances d'annotation sur la bases des mêmes corpus testes. Ceci est expliqué par un meilleur prétraitement des bases d'annotation (issue des ontologies) et surtout par la prise en considération de la notion de « pluriel » dans OntoContext. La conception d'OntoContext avec un module d'annotation permet aussi d'outrepassé les limitations liée à notre définition du contexte. En personnalisant le corpus par le choix précis de *MeSH terms* précis comme requête PubMed notamment, nous avons pu cibler un corpus de résumés qui traite du vieillissement du système immunitaire, cette règle pourra être généralisée aussi aux choix des espèces. La biologie du développement propose une définition du contexte plus large en intégrant la notion de microenvironnement extra et intracellulaire (Orgogozo et al. 2015) ce qui constitue un des point d'amélioration du package.

La deuxième étape notre approche consiste à identifier les voies biologiques mis en jeux d'un contexte à un autre. Dans le chapitre 2 l'utilisation de l'outil PNTHER pour l'analyse des annotations des signatures de gènes mis en jeux démontre que plusieurs processus peuvent être mis en jeux. Aussi le but derrière toutes expériences d'étude du transcriptomes est de voir quels

sont les processus génétique altérés ou activés et quels sont les gènes à l'origine de cette altération. Ceci passe par la modélisation des voies biologiques mise en jeux. La revue de la littérature révèle certaines limites quant à la disponibilité de bases de données de réseaux fonctionnels de gènes. A titre d'exemple la base de données KEGG répertorie moins de 200 voies biologiques (Kanehisa and Goto 2000). Aussi, les outils classiques d'annotation des données d'expression tels que DAVID (Dennis Jr et al. 2003) ont été développés sur la base de sources qui ne tiennent pas compte d'approches de contextualisation telles que Gene Ontology. Pour ces raisons nous avons développé notre propre outil pour la modélisation des voies biologique. Nous avons commencé par définir une voie biologique. Cette étape est cruciale car elle va délémiter les sources d'information pour la modélisation. Nous avons conclue qu'une voie biologique est un ensemble de gènes proches sémantiquement (sur la base de l'ontologie *Biological Proces* de GO) qui interagissent ensemble pour assurer la ou les mêmes fonctions biologiques. Nous avons développé alors, une méthode de modélisation basée sur la proximité sémantique et les données d'interactions à partir des bases de données publiques. Nous avons démontré ici que l'apport des données d'interactions est important pour la classification et le regroupement des gènes en voies biologiques. Dans littérature le modèle proposée par Guo et coll. (Guo et al. 2006) permet uniquement l'identification de couple de protéines susceptibles d'intervenir dans la même voie biologique mais pas de modéliser les voies. Nous avons même testé une approche de calcule de similarité sémantique et le regroupement et le résultat obtenu ne permettait pas de retrouver nos voies biologiques KEGG de départ. Alors que la combinaison des données de proximité et d'interaction avec un algorithme génétique pour la classification nous a permis de regrouper les gènes par voies biologiques. D'autres tests de validation qui prennent en compte un nombre plus importants de gènes apparentant à des voies biologiques KEGG plus diversifiées sont en court, pour l'optimisation de la fonction *fitness* et la diminution du taux d'erreur, mais aussi pour évaluer la spécificité et la sensibilité de la méthode. A ce point du développement nous avons pu uniquement développer une méthode de regroupement des gènes par voies biologiques. Nous travaillerons sur le développement d'une plateforme pour la visualisation des modèles des voies biologiques, ultérieurement. Par rapport à d'autres outils de référence tel qu'IPA qui utilise une approche basée sur le criblage semi-automatique de la littérature, l'approche proposé ici est une approche automatique qui demande moins de moyens humains et matériels. Nous travaillons aussi sur la publication d'un package Python pour faciliter encore plus l'utilisation. Dans la base de données LINCS (Vempati et al. 2014), les réseaux de gènes sont modélisés en fonction du profil d'expression des gènes dans les cellules,

mais cette initiative fait intervenir un consortium de laboratoires avec des moyens expérimentaux considérables.

D'un point de vue global l'approche présentée ici nous a permis à partir d'un corpus de résumées d'articles de générer une base de connaissance sous la forme d'un ensemble de signatures de gènes contextualisées, et regroupé en voies biologiques. Cette base de connaissance constitue aussi une source d'annotation exhaustive et personnalisable en fonction du sujet d'étude. En effet le corpus de textes peut être modifié soit par une requête dans la base de données PubMed soit par l'utilisation d'un corpus de textes personnels. Le dernier argument est un argument économique, en effet l'ensemble des outils développés au sein de ce projet sont des outils qui sont publiés en ligne gratuitement, les ressources utilisées sont aussi des ressources disponibles gratuitement. Un des objectifs de la thèse est de produire une base de connaissances sur le vieillissement du système immunitaire. A ce point de développement nous avons produit une base de signatures par le criblage de la littérature, elle contient plus de 2000 concepts cellulaires, 400 anatomiques et 3000 pathologiques. Ceci permet à un biologiste d'avoir une information exhaustive sur les populations cellulaires atteinte au cours du vieillissement du système immunitaire, en fonction des compartiments anatomiques et surtout les pathologies associées avec une liste des gènes pour chaque contexte susceptible d'être impliqués dans les processus de vieillissement. Il est indispensable de finaliser la seconde étape pour pouvoir confronter notre démarche aux données d'expression pour pouvoir commencer à annoter les signatures d'expression des gènes issue des puces produite pour le vieillissement.

Conclusion

Nous avons présenté ici une démarche inspirée des méthodologies des sciences de l'information et des sciences des données pour l'analyse et l'interprétation des données biologiques. Nous avons développé ainsi la notion de contextualisation qui en sciences de l'information est liée au sens et dans notre projet elle est liée à l'expression des fonctions biologiques. Notre approche est une approche à deux étapes, la contextualisation des gènes et de leurs produits, puis la modélisation des réseaux de gènes par algorithme génétique. La démarche vise à produire une base de connaissances à partir d'un corpus textuel dans lequel on va ensuite identifier des noms de gènes et de leurs produits co-cités dans un contexte particulier (population cellulaire, compartiment anatomique et pathologie) et décrire les processus de régulation auxquels les gènes sont associés par modélisation des voies biologiques impliquées. Pour cette étape, nous

avons développé un package Python. OntoContext qui a été utilisé pour l'annotation automatique des textes, pour la production des contextes. Nous avons testé OntoContext sur des corpus pré-annotés et nous avons démontré qu'il a de meilleures performances que NCBO Annotator, un outil de référence pour l'annotation textuelle à partir des ontologies biomédicales. Nous avons utilisé OntoContext pour annoter un corpus de textes spécialisés dans le vieillissement du système immunitaire et nous avons construit une base de signatures de gènes contextualisées. OntoContext est actuellement disponible en libre accès sur github : https://github.com/walidbedhiafi/OntoContext1 et sur le répertoire pypi : https://pypi.Python.org/pypi/OntoContext. Dans sa version actuelle OntoContext intègre trois ontologies pour l'annotation textuelle, selon notre définition du contexte. Les points d'amélioration porteront sur l'intégration d'un module de prétraitement des ontologies, pour permettre d'intégrer n'importe quelle ontologie pour l'annotation. En effet, l'intégration d'une ontologie telle que l'ontologie *Human Phenotype Ontology* semble pertinente car cela permettrait d'intégrer les anomalies phénotypiques.

Pour le package de modélisation des voies biologiques, nous avons utilisé une approche basée sur l'apprentissage automatique. Nous avons intégré deux sources d'informations : une ontologie biologique et une base de données d'interactions. Nous avons validé l'outil sur des données connues à partir de la base de données KEGG (des voies biologiques déjà décrites dont les données ont été dissociées pour voir si on peut les reconstruire). Nous avons trouvé que l'outil était capable de prédire des voies biologiques avec un taux d'erreur de 46% en moyenne. Pour un jeu de données de 30 gènes appartenant à 3 vois biologiques intracellulaires différentes. Néanmoins nous voulons valider la méthode sur un nombre plus important de voies biologiques avant de généraliser la méthode sur des groupes de gènes sans aucune connaissance préalable. Des étapes de validation et d'optimisation de cet outil sont en cours avant la publication du package. Cette seconde étape est indispensable pour pouvoir confronter les bases d'annotation complètes pour l'étude du transcriptome du vieillissement du système immunitaire. A ce point du projet notre revue de la littérature ne nous nous a pas permis d'identifier un outil similaire pour la comparaison.

Globalement, cette approche de contextualisation et de modélisation de réseaux de gènes permet de générer automatiquement une base de connaissance sur un domaine d'étude bien particulier. L'annotation d'un corpus de textes particulier et la modélisation des voies biologiques permet de situer les gènes et leurs produits dans un compartiment anatomique (organe, tissus…), dans une population cellulaire mais aussi de décrire les mécanismes de

régulations qui en découlent (voies biologiques) et les pathologies associées (contexte pathologique). Cette base de connaissances constituera un outil d'aide à l'interprétation des résultats biologiques. Nous commencerons par appliquer la méthode sur les données générées pour l'analyse du vieillissement du système immunitaire.

Annexe 1: Table récapitulative des cent concepts cellulaires les plus cités les dans le corpus de résumés sur le vieillissement du système immunitaire

Concept cellulaire identifié	Nombre de citations
CELL	77 106
T CELL	36 670
LYMPHOCYTE	29 675
MONONUCLEAR CELL	13 848
B CELL	13 340
LEUKOCYTE	12 360
PERIPHERAL BLOOD MONONUCLEAR CELL	9 643
MONOCYTE	9 362
NEUTROPHIL	9 355
MACROPHAGE	8 642
NATURAL KILLER CELL	5 396
EOSINOPHIL	4 668
GRANULOCYTE	4 580
DENDRITIC CELL	4 047
DENDRITIC CELL, HUMAN	4 016
BLOOD CELL	3 425
STEM CELL	3 417
REGULATORY T CELL	3 053
ENDOTHELIAL CELL	2 738
PLATELET	2 710
CD4-POSITIVE, CD25-POSITIVE, ALPHA-BETA REGULATORY T CELL	2 638
CD8-POSITIVE, ALPHA-BETA CYTOTOXIC T CELL	2 416
PLASMA CELL	2 357
CYTOTOXIC T CELL	2 248
MATURE NEUTROPHIL	2 218
BASOPHIL	2 044
ERYTHROCYTE	1 965
EPITHELIAL CELL	1 777
MAST CELL	1 766
INFLAMMATORY CELL	1 524
FIBROBLAST	1 486
HELPER T CELL	1 248
THYMOCYTE	1 194
MATURE EOSINOPHIL	1 170
MATURE BASOPHIL	1 149
HEMATOPOIETIC STEM CELL	1 064
MEMORY T CELL	1 061
ALVEOLAR MACROPHAGE	1 031
LANGERHANS CELL	878
PHAGOCYTE	874
T-HELPER 17 CELL	759
MEMORY B CELL	715
PLASMACYTOID DENDRITIC CELL	709
BONE MARROW CELL	660

HEMATOPOIETIC CELL	659
STROMAL CELL	628
T-HELPER 2 CELL	620
MUSCLE CELL	614
PROFESSIONAL ANTIGEN PRESENTING CELL	612
T-HELPER 1 CELL	583
ENUCLEATE ERYTHROCYTE	557
NUCLEATE ERYTHROCYTE	557
KERATINOCYTE	547
MYELOID DENDRITIC CELL	545
CD8-POSITIVE, ALPHA-BETA REGULATORY T CELL	539
EFFECTOR T CELL	509
MYELOID CELL	489
PERIPHERAL BLOOD STEM CELL	483
NAIVE T CELL	483
HEPATOCYTE	468
PLASMACYTOID DENDRITIC CELL, HUMAN	460
NON-TERMINALLY DIFFERENTIATED CELL	441
GRANULOCYTE MONOCYTE PROGENITOR CELL	420
MONONUCLEAR PHAGOCYTE	374
MATURE NATURAL KILLER CELL	367
INTRAEPITHELIAL LYMPHOCYTE	361
NEURON	354
PRECURSOR CELL	349
MESENCHYMAL STEM CELL	348
SMOOTH MUSCLE CELL	344
CD8-ALPHA-BETA-POSITIVE, ALPHA-BETA INTRAEPITHELIAL T CELL	331
GAMMA-DELTA INTRAEPITHELIAL T CELL	331
CD4-POSITIVE, ALPHA-BETA INTRAEPITHELIAL T CELL	331
CD4-NEGATIVE CD8-NEGATIVE GAMMA-DELTA INTRAEPITHELIAL T CELL	331
ALPHA-BETA INTRAEPITHELIAL T CELL	331
CD8-ALPHA ALPHA POSITIVE, GAMMA-DELTA INTRAEPITHELIAL T CELL	331
MEGAKARYOCYTE	326
ERYTHROID PROGENITOR CELL	310
MICROGLIAL CELL	303
CONVENTIONAL DENDRITIC CELL	303
OSTEOCLAST	292
MATURE T CELL	288
ERYTHROID PROGENITOR CELL, MAMMALIAN	285
FOLLICULAR DENDRITIC CELL	281
MATURE B CELL	269
MYELOID DENDRITIC CELL, HUMAN	263
PERITONEAL MACROPHAGE	262

CULTURED CELL	262
SPLENOCYTE	261
GAMMA-DELTA T CELL	223
EPITHELIAL CELL OF THYMUS	223
EPITHELIOID MACROPHAGE	222
FAT CELL	221
FOAM CELL	220
VEIN ENDOTHELIAL CELL	214
ASTROCYTE	212
RETICULOCYTE	197
PIGMENTED CILIARY EPITHELIAL CELL	194
PRO-B CELL	190

Annexe 2: Table récapitulative des cent concepts anatomiques les plus cites dans le corpus de résumés sur le vieillissement du système immunitaire

Concept anatomique identifié	Nombre de citation
BLOOD	42088
TISSUE	13177
ANATOMICAL SYSTEM	10612
LYMPH NODE	5566
BODILY SECRETION	4958
ORGAN	3214
DEATH STAGE	3025
BREAST	2495
HEART	2272
MUCOSA	2093
BRAIN	1943
EPITHELIUM	1445
NERVOUS SYSTEM	1232
NECK	1231
CENTRAL NERVOUS SYSTEM	981
HEAD	956
DIGESTIVE SYSTEM	881
SYNOVIAL FLUID	788
MIDGUT	782
LAMINA	777
LAMINA PROPRIA	693
OBSOLETE VISCERAL MUSCLE	683
ANATOMICAL SPACE	680
EYE	652
NERVE	602
INTESTINE	536
SKELETAL JOINT	457
AXILLARY LYMPH NODE	382
DUCT	375
SALIVA-SECRETING GLAND	370
STOMACH	318
ADIPOSE TISSUE	312
LYMPHOID FOLLICLE	277
TUBE	272
LEG	247
AORTA	227
MEDULLA OF ORGAN	200
LIFE CYCLE	199
ORAL CAVITY	195
ESOPHAGUS	186
NOSE	175
SOLID COMPOUND ORGAN	168
TROPHOBLAST	163
COLONIC MUCOSA	162
SUBMUCOSA	159

THROAT	155
EMBRYO	153
BLOOD BRAIN BARRIER	151
RECTUM	135
BRONCHIAL MUCOSA	134
RENAL SYSTEM	129
CORNEA	120
RETICULOENDOTHELIAL SYSTEM	118
TESTIS	114
MOUTH	113
RETINA	112
ISLET OF LANGERHANS	111
SCALP	102
CELL LAYER	101
ABDOMEN	100
SWEAT	96
DUODENAL MUCOSA	90
PONS	87
UTERINE CERVIX	75
LIGAMENT	74
LARGE INTESTINE	72
WING	68
SYMPATHETIC NERVOUS SYSTEM	66
PERIPHERAL NERVOUS SYSTEM	65
AMNIOTIC FLUID	63
FEMUR	63
PLEURA	62
CELL PART	60
INTERSTITIAL FLUID	57
NEURAL GLOMERULUS	56
CEREBRAL CORTEX	55
ENDOCRINE SYSTEM	53
JEJUNAL MUCOSA	52
TENDON	52
GANGLION	51
VAGINA	51
PITUITARY GLAND	50
RESPIRATORY SYSTEM	49
MAMMALIAN VULVA	48
RED BONE MARROW	47
SEMINAL VESICLE	46
SENSORY NERVE	45
STERNUM	43
UROTHELIUM	41
CYTOTROPHOBLAST	40
EXTRAVILLOUS TROPHOBLAST	39

SYNCYTIOTROPHOBLAST	38
DIAPHRAGM	37
SUBARACHNOID SPACE	37
VENTRAL PANCREATIC DUCT	36
TIBIA	36
ARTERY WALL	35
COLONIC EPITHELIUM	34
MYELIN SHEATH	34
CIRCULATORY SYSTEM	33

Annexe 3: Table récapitulative des cent concepts pathologiques les plus cités dans le corpus de résumés sur le vieillissement du système immunitaire

Concept pathologique identifié	Nombre de citation
CANCER	14468
LYMPHOMA	5749
CARCINOMA	4832
ASTHMA	3831
DISEASE BY INFECTIOUS AGENT	3328
HEPATITIS	3028
ANEMIA	2183
BREAST CANCER	1915
MULTIPLE SCLEROSIS	1850
MELANOMA	1376
HYPERSENSITIVITY REACTION TYPE I DISEASE	1317
DERMATITIS	1303
HEPATITIS C	1249
HEPATITIS B	1166
CHRONIC LYMPHOCYTIC LEUKEMIA	1048
CARDIOMYOPATHY	990
SQUAMOUS CELL CARCINOMA	931
LUNG CANCER	930
ATOPIC DERMATITIS	928
ADENOCARCINOMA	764
SARCOIDOSIS	760
HYPERTENSION	750
TETANUS	676
CHRONIC LEUKEMIA	670
MALARIA	623
SARCOMA	597
ACUTE LEUKEMIA	596
KIDNEY FAILURE	518
PURPURA	466
BREAST CARCINOMA	456
CHRONIC OBSTRUCTIVE PULMONARY DISEASE	446
CORONARY ARTERY DISEASE	445
PROSTATE CANCER	438
DIARRHEA	426
HISTIOCYTOSIS	400
COMMON VARIABLE IMMUNODEFICIENCY	377
FUNGAL INFECTIOUS DISEASE	356
ADULT T-CELL LEUKEMIA	353
SKIN DISEASE	352
EXANTHEM	347
NEPHROTIC SYNDROME	340
CYSTIC FIBROSIS	332
PANCYTOPENIA	330
OVARIAN CANCER	316

PERIODONTAL DISEASE	282
PULMONARY FIBROSIS	275
MASTOCYTOSIS	275
RELAPSING-REMITTING MULTIPLE SCLEROSIS	266
VIRAL HEPATITIS	260
CELIAC DISEASE	258
FOOD ALLERGY	250
PRIMARY BILIARY CIRRHOSIS	246
CHRONIC GRANULOMATOUS DISEASE	245
PULMONARY SARCOIDOSIS	233
THROMBOCYTOSIS	212
NECK CANCER	210
HEAD AND NECK CANCER	209
VASCULAR DISEASE	207
PERTUSSIS	205
PANCREATIC CANCER	205
IDIOPATHIC PULMONARY FIBROSIS	201
KAWASAKI DISEASE	201
TONSILLITIS	195
HEMOPHAGOCYTIC LYMPHOHISTIOCYTOSIS	194
MARGINAL ZONE B-CELL LYMPHOMA	191
INTERSTITIAL LUNG DISEASE	191
HYPERGAMMAGLOBULINEMIA	189
HYPERCHOLESTEROLEMIA	184
RETINITIS PIGMENTOSA	181
SYSTEMIC MASTOCYTOSIS	181
HEMOPHILIA	170
CONTACT DERMATITIS	170
ACUTE PROMYELOCYTIC LEUKEMIA	168
IDIOPATHIC INTERSTITIAL PNEUMONIA	168
CANDIDIASIS	159
ESSENTIAL HYPERTENSION	158
PLASMACYTOMA	154
ORAL SQUAMOUS CELL CARCINOMA	153
PRIMARY CUTANEOUS AMYLOIDOSIS	148
LYMPHADENITIS	148
POLYNEUROPATHY	147
GOITER	144
PLASMODIUM FALCIPARUM MALARIA	144
DIPHTHERIA	143
DEMYELINATING DISEASE	142
DERMATOMYOSITIS	141
AGGRESSIVE PERIODONTITIS	139
GINGIVITIS	139
AORTIC ANEURYSM	135
CRYOGLOBULINEMIA	130

HYPOTHYROIDISM	130
SPINAL CORD DISEASE	129
AGAMMAGLOBULINEMIA	128
AUTOIMMUNE HEPATITIS	123
THYROID CANCER	123
DISEASE OF METABOLISM	122
HELICOBACTER PYLORI INFECTIOUS DISEASE	121
MALT LYMPHOMA	121
CHOLERA	121

Annexe 4: Table récapitulative des cent concepts protéiques les plus cités dans le corpus de résumés sur le vieillissement du système immunitaire

Concept protéique identifié	Nombre de citations
cytokine	1156
IL2	827
IL-2	773
IFNgamma	517
IFN-gamma	491
CD4	481
TNFalpha	397
IgE	395
IL6	393
IL-6	388
TNF-alpha	363
IL10	330
IL4	329
IL-10	320
IL-4	312
GCSF	293
G-CSF	289
GMCSF	278
IgG	278
GM-CSF	274
CD8	256
antibody	253
IgA	248
IgM	213
IL-1	205
IL1	205
IL8	185
IL12	185
IL-12	182
IL-8	180
IFN	177
TNF	161
immunoglobulin	160
HLADR	160
TCR	153
autoantibody	152
IL5	151
IL-5	147
IFNalpha	145
HLA-DR	143
IFN-alpha	141
PHA	141
monoclonal antibody	131
adhesion molecule	131

ICAM1	129
ICAM-1	127
proinflammatory cytokine	125
CSF	122
lymphokine	121
CD3	115
interferon	113
CD34	109
RA	98
TGFbeta	95
Ig	94
IL2R	94
IL-1 beta	94
mitogen	91
TGF-beta	91
IL1 beta	88
TNF alpha	81
p53	80
immune complex	79
LDL	78
NFkappaB	78
IL3	77
chemokine	76
rIL2	76
ECP	76
IL-3	75
rIL-2	75
immunoregulatory	73
VEGF	73
CNS	73
CD28	72
NF-kappaB	72
inflammatory cytokine	71
epitope	71
IL-2R	70
Fas	70
CD56	69
insulin	69
interleukin-2	68
Th2 cytokine	68
TIL	67
interleukin2	67
CD14	67
enzyme	65
IL13	65
CD25	64

rhGCSF	63
IL-13	63
PWM	62
rhG-CSF	62
interferon-gamma	61
mRNA	61
CD16	60
IL7	60
CD5	59

Annexe 5: Table récapitulative des cent concepts ARN les plus cités dans le corpus de résumés sur le vieillissement du système immunitaire

Concept ARN identifié	Nombre de citations
cytokine	1 156
IL2	827
IL-2	773
IFNgamma	517
IFN-gamma	491
CD4	481
TNFalpha	397
IgE	395
IL6	393
IL-6	388
TNF-alpha	363
IL10	330
IL4	329
IL-10	320
IL-4	312
GMCSF	278
GM-CSF	274
IL-1	205
IL1	205
IL8	185
IL12	185
IL-12	182
IL-8	180
IL5	151
IL-5	147
CSF	122
interferon	113
IL2R	94
IL-1 beta	94
Ig	94
IL1 beta	88
NFkappaB	78
VEGF	73
NF-kappaB	72
IL-2R	70
IL13	65
IL-13	63
mRNA	61
IL15	59
IL-15	57
CD69	51
IL17	46
IL-17	45
CD40L	40
GR	38

autoantigen	35
CD38	34
CD95	27
PD-1	23
PSA	22
PD1	22
HGF	22
c-myc	19
IPF	18
SM	14
cytokine mRNA	14
RANKL	12
IL-9	12
IL9	12
CD36	12
TPO	11
RP	11
IL-4 mRNA	10
TGFalpha	9
HPS	9
TGF-alpha	8
HCV-RNA	8
HCVRNA	8
HCV RNA	8
protein tyrosine kinase	7
p55	7
IL4 mRNA	7
IL-2 mRNA	7
IL2 mRNA	7
IL-10 mRNA	7
IL10 mRNA	7
short arm	6
INS	6
IL-5 mRNA	6
IL5 mRNA	6
G-CSFR	6
GCSFR	6
CHC	6
TGF beta	5
LFA-3	5
LFA3	5
IL-8 mRNA	5
IL8 mRNA	5
Foxp3 mRNA	5
type-2 cytokine	4
Tax	4

IL-6 mRNA	4
IL6 mRNA	4
TTP	3
TNF mRNA	3
TF mRNA	3
S1	3
RNA	3
MGF	3

Annexe 6: Table récapitulative des cent concepts ADN les plus cités dans le corpus de résumés sur le vieillissement du système immunitaire

Concept ADN identifié	Nombre de citations
detoxification gene	1 156
autosomal dominant gene	827
mature hematopoietic element	773
beta 2binde site	517
beta 2-binding site	491
WT1	481
ICU	397
OLR1 gene	395
OME	393
chromosome rearrangement	388
lymphokine gene	363
glutathioneStransferase M1 mutant allele	330
IgE	329
glutathione-S-transferase M1 mutant allele	320
HLA antibody	312
antigranulocyte antibody	293
LCT	289
Tcellspecific transact regulatory factor	278
T-cell-specific trans-acting regulatory factor	278
8q24	274
light chain gene	256
growth-regulated gene	253
tumor suppressor gene	248
mcr and far3'MBR region	213
mcr and far3'-MBR region	185
MBR	185
CTL gene	182
lymphoid organ and/or other mucosal site	180
VH gene	177
light chain	161
fibril protein	160
C-terminal end	153
lambda-light-chain fragment	152
t (14 ; 18)	147
11q	143
14q	131
6q	131
20q	122
hepatic site	115
Y chromosome	109
IFN-gamma gene	98
IFNgamma gene	95

notyetidentified gene	94
not-yet-identified gene	91
chromosome 9	80
FHL 1 locus	79
chromosome 1	78
RB gene	77
regulatory element	76
B cell repertoire	75
chromosome 3q27	75
serum factor	73
C3b receptor	73
complement fragment	72
HIV PI	71
MTHFR gene	71
ALD gene	69
ALDP	67
VLCFA	65
viral DNA	63
hst1 gene	61
hst-1 gene	60
human UbB gene	60
DEB	59
GSTT1 gene	58
ribosomal gene	57
phVEGF165 gene	57
RNA transcript	56
betathalassaemia allele	56
beta-thalassaemia allele	54
BCG inoculation site	54
GM1	54
inflammatory site	53
gamma detection probe	53
polymorphism	53
13q14	51
exon b3	51
recombination mutant	50
HLA-A locus	49
HLAA locus	49
p16 gene	48
cCbl and Cblb gene	48
c-Cbl	48
Cbl-b gene	45
HLA gene	44
preBrelated gene	44
pre-B-related gene	44
DRD3 gene	44

polymorphic dinucleotide	42
CAG repeat	42
trinucleotide repeat	42
c-fgr	39
chromosomal locus	38
chromosome 7	37
T-cell receptor beta chain	37
Tcell receptor beta chain	36
perforin promoter	35
perforin	34
mRNA	32

Annexe 7: Table récapitulative des signatures de gènes identifiées pour les 4 contextes à partir du corpus sur le vieillissement du système immunitaire

Contexte	Noms identifiés à partir des résumées	Symboles et noms usuels
BCELL/BLOOD/ CANCER	CD5	T-cell surface glycoprotein CD5;CD5;ortholog
	IL-2,IL2	Interleukin-2;IL2;ortholog
	IL4	Interleukin-4;IL4;ortholog
	LAIR-1,LAIR1	Leukocyte-associated immunoglobulin-like receptor 1;LAIR1;ortholog
	CdA	Cytidine deaminase;CDA;ortholog
	TNF-alpha	Tumor necrosis factor;TNF;ortholog
	MAGE1	Melanoma-associated antigen E1;MAGEE1;ortholog
	Tc1	Uncharacterized protein C8orf4;C8orf4;ortholog
	IL6,IL-6	Interleukin-6;IL6;ortholog
	CD2	T-cell surface antigen CD2;CD2;ortholog
	IL7,IL-7	Interleukin-7;IL7;ortholog
	CCND3	G1/S-specific cyclin-D3;CCND3;ortholog
	VEGFC,VEGF-C	Vascular endothelial growth factor C;VEGFC;ortholog
	Pgp	Phosphoglycolate phosphatase;PGP;ortholog
	CD40L	CD40 ligand;CD40LG;ortholog
	ZAP70	Tyrosine-protein kinase ZAP-70;ZAP70;ortholog
	EPO	Erythropoietin;EPO;ortholog
	MAGE1	Melanoma-associated antigen 1;MAGEA1;ortholog
	AITR	Tumor necrosis factor receptor superfamily member 18;TNFRSF18;ortholog
	EPO	Eosinophil peroxidase;EPX;ortholog
	Ph1	Polyhomeotic-like protein 1;PHC1;ortholog
	bcl2	Apoptosis regulator Bcl-2;BCL2;ortholog
	LFA3	Lymphocyte function-associated antigen 3;CD58;ortholog
	CD20	B-lymphocyte antigen CD20;MS4A1;ortholog
	beta-galactosidase	Beta-galactosidase;GLB1;ortholog
	gp120	Inter-alpha-trypsin inhibitor heavy chain H4;ITIH4;ortholog
	Cdc7	Cell division cycle 7-related protein kinase;CDC7;ortholog
	PRELI	PRELI domain-containing protein 1, mitochondrial;PRELID1;ortholog
	IL10,IL-10	Interleukin-10;IL10;ortholog

	IgA	B-cell antigen receptor complex-associated protein alpha chain;CD79A;ortholog
T CELL, BLOOD, CANCER	CD11a	Cyclin-dependent kinase 11A;CDK11A;ortholog
	IRF4	Interferon regulatory factor 4;IRF4;ortholog
	interleukin-2,TCGF,IL-2,IL2	Interleukin-2;IL2;ortholog
	IL4,IL-4	Interleukin-4;IL4;ortholog
	MAGE3	Melanoma-associated antigen 3;MAGEA3;ortholog
	LAIR-1,LAIR1	Leukocyte-associated immunoglobulin-like receptor 1;LAIR1;ortholog
	DPP	Dentin sialophosphoprotein;DSPP;ortholog
	PBT	Mast/stem cell growth factor receptor Kit;KIT;ortholog
	CLC	Galectin-10;CLC;ortholog
	TNF-alpha	Tumor necrosis factor;TNF;ortholog
	CAP1,CAP-1	TNF receptor-associated factor 3;TRAF3;ortholog
	IL15,IL-15	Interleukin-15;IL15;ortholog
	MAGE1	Melanoma-associated antigen E1;MAGEE1;ortholog
	SpS	Selenide, water dikinase 1;SEPHS1;ortholog
	STAT1	Signal transducer and activator of transcription 1-alpha/beta;STAT1;ortholog
	Tc1	Uncharacterized protein C8orf4;C8orf4;ortholog
	TIL	Toll-like receptor 1;TLR1;ortholog
	IL6,IL-6	Interleukin-6;IL6;ortholog
	CD26	Dipeptidyl peptidase 4;DPP4;ortholog
	CD1a	T-cell surface glycoprotein CD1a;CD1A;ortholog
	PHC	Phosphate carrier protein, mitochondrial;SLC25A3;ortholog
	Stat5,STAT5	Signal transducer and activator of transcription 5A;STAT5A;ortholog
	CRF	Corticoliberin;CRH;ortholog
	CD11a	Integrin alpha-L;ITGAL;ortholog
	p55	DNA polymerase subunit gamma-2, mitochondrial;POLG2;ortholog
	CD2	T-cell surface antigen CD2;CD2;ortholog
	CD6	T-cell differentiation antigen CD6;CD6;ortholog
	PSA	Phosphoserine aminotransferase;PSAT1;ortholog
	p55	Protein disulfide-isomerase;P4HB;ortholog
	Ksp37	Fibroblast growth factor-binding protein 2;FGFBP2;ortholog

IL7,IL-7	Interleukin-7;IL7;ortholog
PSA	Prostate-specific antigen;KLK3;ortholog
CD7	T-cell antigen CD7;CD7;ortholog
CEA	Carcinoembryonic antigen-related cell adhesion molecule 5;CEACAM5;ortholog
Sp17	Sperm surface protein Sp17;SPA17;ortholog
BAK	Bcl-2 homologous antagonist/killer;BAK1;ortholog
Pgp	Phosphoglycolate phosphatase;PGP;ortholog
CD40L	CD40 ligand;CD40LG;ortholog
EPO	Erythropoietin;EPO;ortholog
MAGE1	Melanoma-associated antigen 1;MAGEA1;ortholog
CD14	Monocyte differentiation antigen CD14;CD14;ortholog
Ep-CAM,EpCAM	Epithelial cell adhesion molecule;EPCAM;ortholog
DDP	Mitochondrial import inner membrane translocase subunit Tim8 A;TIMM8A;ortholog
AITR	Tumor necrosis factor receptor superfamily member 18;TNFRSF18;ortholog
EPO	Eosinophil peroxidase;EPX;ortholog
PPD	4-hydroxyphenylpyruvate dioxygenase;HPD;ortholog
B2M	Beta-2-microglobulin;B2M;ortholog
IFNg,IFN-gamma	Interferon gamma;IFNG;ortholog
bcl2	Apoptosis regulator Bcl-2;BCL2;ortholog
LFA3	Lymphocyte function-associated antigen 3;CD58;ortholog
Tat	Tyrosine aminotransferase;TAT;ortholog
beta-galactosidase	Beta-galactosidase;GLB1;ortholog
TLR9	Toll-like receptor 9;TLR9;ortholog
RelA	Transcription factor p65;RELA;ortholog
NKG2A	NKG2-A/NKG2-B type II integral membrane protein;KLRC1;ortholog
p55	55 kDa erythrocyte membrane protein;MPP1;ortholog
gp120	Inter-alpha-trypsin inhibitor heavy chain H4;ITIH4;ortholog
CTLA4	Cytotoxic T-lymphocyte protein 4;CTLA4;ortholog
CARS	Cysteine--tRNA ligase, cytoplasmic;CARS;ortholog
Flt3L	Fms-related tyrosine kinase 3 ligand;FLT3LG;ortholog

COX2	Cytochrome c oxidase subunit 2;MT-CO2;ortholog
LAK	Alpha-protein kinase 1;ALPK1;ortholog
p43	Elongation factor Tu, mitochondrial;TUFM;ortholog
COX-2,COX2	Prostaglandin G/H synthase 2;PTGS2;ortholog
Ksp37	Transmembrane protein 72;TMEM72;ortholog
IL10,IL-10	Interleukin-10;IL10;ortholog
GM-CSF,GMCSF	Granulocyte-macrophage colony-stimulating factor;CSF2;ortholog
RelB	Transcription factor RelB;RELB;ortholog
HLA-C	HLA class I histocompatibility antigen, Cw-17 alpha chain;HLA-C;ortholog
HER2	Receptor tyrosine-protein kinase erbB-2;ERBB2;ortholog
IgA	B-cell antigen receptor complex-associated protein alpha chain;CD79A;ortholog
ACTL	Acetyl-CoA acetyltransferase, cytosolic;ACAT2;ortholog
Foxp3	Forkhead box protein P3;FOXP3;ortholog
ADA	Adenosine deaminase;ADA;ortholog
Tg	Thyroglobulin;TG;ortholog
Tim3,Tim-3	Hepatitis A virus cellular receptor 2;HAVCR2;ortholog
CRF	C1q-related factor;C1QL1;ortholog
IL16	Pro-interleukin-16;IL16;ortholog
CD19	B-lymphocyte antigen CD19;CD19;ortholog
PPD	Protein argonaute-2;AGO2;ortholog
ATK	Tyrosine-protein kinase BTK;BTK;ortholog
PSK	Seizure 6-like protein 2;SEZ6L2;ortholog
BRM	Probable global transcription activator SNF2L2;SMARCA2;ortholog
CD34	Hematopoietic progenitor cell antigen CD34;CD34;ortholog
CD28	T-cell-specific surface glycoprotein CD28;CD28;ortholog
CLC	Cardiotrophin-like cytokine factor 1;CLCF1;ortholog
GA	Glutaminase liver isoform, mitochondrial;GLS2;ortholog
p55	Fascin;FSCN1;ortholog
Tc2	Transcobalamin-2;TCN2;ortholog
TDL	Apelin receptor early endogenous ligand;APELA;ortholog
IL-18,IL18	Interleukin-18;IL18;ortholog

	APR	Prolow-density lipoprotein receptor-related protein 1;LRP1;ortholog
	IL18R	Interleukin-18 receptor 1;IL18R1;ortholog
	UCHL1	Ubiquitin carboxyl-terminal hydrolase isozyme L1;UCHL1;ortholog
	G-CSF,GCSF	Granulocyte colony-stimulating factor;CSF3;ortholog
	CCR7	C-C chemokine receptor type 7;CCR7;ortholog
	MIC2	CD99 antigen;CD99;ortholog
	p55	Interleukin-2 receptor subunit alpha;IL2RA;ortholog
	CD38	ADP-ribosyl cyclase/cyclic ADP-ribose hydrolase 1;CD38;ortholog
	CAP1	Adenylyl cyclase-associated protein 1;CAP1;ortholog
	CD69	Early activation antigen CD69;CD69;ortholog
	AITRL	Tumor necrosis factor ligand superfamily member 18;TNFSF18;ortholog
	APR	Phorbol-12-myristate-13-acetate-induced protein 1;PMAIP1;ortholog
	CD4	T-cell surface glycoprotein CD4;CD4;ortholog
	PSA	Puromycin-sensitive aminopeptidase;NPEPPS;ortholog
	Tc1	Transcobalamin-1;TCN1;ortholog
	CAP1	Prostasin;PRSS8;ortholog
	Tc1	Thiamine transporter 1;SLC19A2;ortholog
T CELL, BLOOD, DERMATITIS	Leu1	T-cell surface glycoprotein CD5;CD5;ortholog
	IL-2,IL2	Interleukin-2;IL2;ortholog
	IL4,IL-4	Interleukin-4;IL4;ortholog
	CD43	Leukosialin;SPN;ortholog
	TNF-alpha	Tumor necrosis factor;TNF;ortholog
	Leu1	Leukemia-associated protein 1;DLEU1;ortholog
	GATA3	Trans-acting T-cell-specific transcription factor GATA-3;GATA3;ortholog
	Tc1	Uncharacterized protein C8orf4;C8orf4;ortholog
	CD26	Dipeptidyl peptidase 4;DPP4;ortholog
	IL5,IL-5	Interleukin-5;IL5;ortholog
	CD30	Tumor necrosis factor receptor superfamily member 8;TNFRSF8;ortholog
	IFN-gamma	Interferon gamma;IFNG;ortholog
	CD86	T-lymphocyte activation antigen CD86;CD86;ortholog
	IL3,IL-3	Interleukin-3;IL3;ortholog

	TCC	Sideroflexin-1;SFXN1;ortholog
	CCR4	Nocturnin;CCRN4L;ortholog
	CCR5	C-C chemokine receptor type 5;CCR5;ortholog
	IL10,IL-10	Interleukin-10;IL10;ortholog
	L-selectin	L-selectin;SELL;ortholog
	IL4R	Interleukin-4 receptor subunit alpha;IL4R;ortholog
	CD40	Tumor necrosis factor receptor superfamily member 5;CD40;ortholog
	CCR4	C-C chemokine receptor type 4;CCR4;ortholog
	CCR3	C-C chemokine receptor type 3;CCR3;ortholog
	CCR4	CCR4-NOT transcription complex subunit 6;CNOT6;ortholog
	IL13,IL-13	Interleukin-13;IL13;ortholog
	Txk	Tyrosine-protein kinase TXK;TXK;ortholog
	IL8	Interleukin-8;CXCL8;ortholog
	CD4	T-cell surface glycoprotein CD4;CD4;ortholog
	SOCS3,SOCS-3	Suppressor of cytokine signaling 3;SOCS3;ortholog
	CXCR3	C-X-C chemokine receptor type 3;CXCR3;ortholog
	CD80	T-lymphocyte activation antigen CD80;CD80;ortholog
	Tc1	Transcobalamin-1;TCN1;ortholog
	Tc1	Thiamine transporter 1;SLC19A2;ortholog
T CELL, LYMPHNODE, CANCER	CD11a	Cyclin-dependent kinase 11A;CDK11A;ortholog
	IL-2,IL2	Interleukin-2;IL2;ortholog
	IL4,IL-4	Interleukin-4;IL4;ortholog
	PTC	Proto-oncogene tyrosine-protein kinase receptor Ret;RET;ortholog
	TIL	Toll-like receptor 1;TLR1;ortholog
	FasL	Tumor necrosis factor ligand superfamily member 6;FASLG;ortholog
	CD11a	Integrin alpha-L;ITGAL;ortholog
	PTC	Taste receptor type 2 member 38;TAS2R38;ortholog
	RLNL	Insulin-like 3;INSL3;ortholog
	CD7	T-cell antigen CD7;CD7;ortholog
	PR	Progesterone receptor;PGR;ortholog
	BCL1	G1/S-specific cyclin-D1;CCND1;ortholog
	IFN-gamma	Interferon gamma;IFNG;ortholog
	BCL6	B-cell lymphoma 6 protein;BCL6;ortholog
	RCAS1	Receptor-binding cancer antigen expressed on SiSo cells;EBAG9;ortholog

PR	Endogenous retrovirus group K member 104 Pro protein;HERV-K104;ortholog
HLA-G	HLA class I histocompatibility antigen, alpha chain G;HLA-G;ortholog
LAK	Alpha-protein kinase 1;ALPK1;ortholog
IL10,IL-10	Interleukin-10;IL10;ortholog
PTC	Coagulation factor IX;F9;ortholog
CD19	B-lymphocyte antigen CD19;CD19;ortholog
BRM	Probable global transcription activator SNF2L2;SMARCA2;ortholog
CD34	Hematopoietic progenitor cell antigen CD34;CD34;ortholog
PR	Endogenous retrovirus group K member 9 Pol protein;ERVK-9;ortholog
Cyclin	Proliferating cell nuclear antigen;PCNA;ortholog
VEGFA	Vascular endothelial growth factor A;VEGFA;ortholog
CD4	T-cell surface glycoprotein CD4;CD4;ortholog
ILT3	Leukocyte immunoglobulin-like receptor subfamily B member 4;LILRB4;ortholog

Annexe 8 : Similarité sémantique entre gènes

Dans GO, un produit de gènes peut être annoté par un ou plusieurs termes. De ce fait, les mesures développées pour la proximité des termes ont été extrapolées pour la mesure de la proximité entre produits de gènes.

➢ La mesure de Wang (Wang et al. 2007) :

A partir de la similarité entre termes, Wang a développé une mesure de similarité gène à gène.

$$Sim(G1, G2) = \frac{\sum_{1\leq i\leq n} Sim(go1i, GO2) + \sum_{1\leq j\leq m} Sim(go2j, GO1)}{n+m}$$

Avec G1 et G2 deux gènes différents annotés simultanément par un set GO1= {go11, go12,…, go1i} et GO2= {go21, go22,…,go2i},

et $Sim(go, GO) = \max(Sim(go, goi)_{1\leq i\leq k}$ la similarité entre un terme et un set de termes.

➢ GS2 (Ruths et al. 2009) :

Ruths et son équipe ont développé une distance hybride plus complexe que la distance de Wang et qui tient compte de la position des annotations dans l'ontologie et des différentes relations entre concepts (annotations). Cette distance présente l'avantage d'être plus spécifique que celle de Wang, son implémentation est plus rapide et surtout elle permet de mesurer les similarités gène à gène ainsi que la distance entre un gène et un groupe de gènes.

Liste des figures

Liste des tables

Références bibliographiques

Aakula, A., Leivonen, S.-K., Hintsanen, P., Aittokallio, T., Ceder, Y., Børresen-Dale, A.-L., Perälä, M., Östling, P., and Kallioniemi, O., 2015. MicroRNA-135b regulates ERα, AR and HIF1AN and affects breast and prostate cancer cell growth. *Molecular Oncology*, 9 (7), 1287–1300.

Abdul Rahman, F., Shamsuddin, S. M., Hassan, S., and Abu Haris, N., 2016. A Review of KDD-Fouille des données Framework and Its Application in Logistics and Transportation. *International Journal of Supply Chain Management*, 5 (2), 77–84.

Agresti, A. and Kateri, M., 2011. Categorical Data Analysis. *In*: Lovric, M., ed. *International Encyclopedia of Statistical Science* [online]. Springer Berlin Heidelberg, 206–208. Available from: http://link.springer.com.gate2.inist.fr/referenceworkentry/10.1007/978-3-642-04898-2_161 [Accessed 4 Jan 2017].

Airoldi, Edoardo M. 2007. « Getting Started in Probabilistic Graphical Models ». *PLOS Computational Biology* 3 (12): e252. doi:10.1371/journal.pcbi.0030252.

Alba, Richard D. 1973. « A graph-theoretic definition of a sociometric clique ». *The Journal of Mathematical Sociology* 3 (1): 113-26. doi:10.1080/0022250X.1973.9989826.

Alpaydin, E., 2010. *Introduction to machine learning*. 2nd ed. Cambridge, Mass: MIT Press.

Anderson, C. S., DeDiego, M. L., Topham, D. J., and Thakar, J., 2016. Boolean Modeling of Cellular and Molecular Pathways Involved in Influenza Infection. *Computational and Mathematical Methods in Medicine* [online], 2016. Available from: https://www-ncbi-nlm-nih-gov.gate2.inist.fr/pmc/articles/PMC4769743/ [Accessed 26 Jan 2017].

Anon., 1996. Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nature Genetics*, 14 (4), 457–460.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G., 2000. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25 (1), 25–29.

Aspray, W., 1997. The Intel 4004 microprocessor: what constituted invention? *IEEE Annals of the History of Computing*, 19 (3), 4–15.

Attwood, T. K., 2000. The Babel of Bioinformatics. *Science*, 290 (5491), 471–473.

Attwood, T. K., Gisel, A., Eriksson, N.-E., and Bongcam-Rudloff, E., 2011. Concepts, Historical Milestones and the Central Place of Bioinformatics in Modern Biology: A European Perspective. [online]. Available from: http://www.intechopen.com/books/bioinformatics-trends-and-methodologies/concepts-historical-milestones-and-the-central-place-of-bioinformatics-in-modern-biology-a-european- [Accessed 4 Jan 2017].

Aumsuwan, P., Khan, S. I., Khan, I. A., Walker, L. A., and Dasmahapatra, A. K., 2016. Gene expression profiling and pathway analysis data in MCF-7 and MDA-MB-231 human breast cancer cell lines treated with dioscin. *Data in Brief*, 8, 272–279.

Auroux, Sylvain. 1998. *La raison, le langage et les normes*. Presses Universitaires de France.

https://halshs.archives-ouvertes.fr/halshs-00529142.

Avery, O. T., MacLeod, C. M., and McCarty, M., 1944. Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types. *Journal of Experimental Medicine*, 79 (2), 137–158.

Bada, M., Eckert, M., Evans, D., Garcia, K., Shipley, K., Sitnikov, D., Baumgartner, W. A., Cohen, K. B., Verspoor, K., Blake, J. A., and Hunter, L. E., 2012. Concept annotation in the CRAFT corpus. *BMC Bioinformatics*, 13, 161.

Bard, J., Rhee, S. Y., and Ashburner, M., 2005. An ontology for cell types. *Genome Biology*, 6 (2), R21.

Bassi, S., 2007. A Primer on Python for Life Science Researchers. *PLOS Computational Biology*, 3 (11), e199.

Bates, M. J. and Maack, M. N., eds., 2009. *Encyclopedia of Library and Information Sciences, Third Edition* [online]. CRC Press. Available from: http://www.crcnetbase.com/doi/book/10.1081/E-ELIS3 [Accessed 26 Dec 2016].

Bayes, M. and Price, M., 1763. An Essay towards Solving a Problem in the Doctrine of Chances. By the Late Rev. Mr. Bayes, F. R. S. Communicated by Mr. Price, in a Letter to John Canton, A. M. F. R. S. *Philosophical Transactions*, 53, 370–418.

Bechhofer, S., 2009. OWL: Web Ontology Language. *In*: LIU, L. and ÖZSU, M. T., eds. *Encyclopedia of Database Systems* [online]. Springer US, 2008–2009. Available from: http://link.springer.com.gate2.inist.fr/referenceworkentry/10.1007/978-0-387-39940-9_1073 [Accessed 3 Jan 2017].

BEN M'Barek, M., 2016. *Développement de modèles d'interactions géniques en se basant sur les algorithmes d'apprentissages automatique*. Institut Supérieur d'Informatique. Stage Master 2.

Berkson, J., 1944. Application of the Logistic Function to Bio-Assay. *Journal of the American Statistical Association*, 39 (227), 357–365.

Berkson, J., 1951. Why I Prefer Logits to Probits. *Biometrics*, 7 (4), 327–339.

Berners-Lee, T., Cailliau, R., Groff, J.-F., and Pollermann, B., 1992. World-Wide Web: The Information Universe. *Electronic networking*, 2 (1), 52–58.

Berners-Lee, T., Hendler, J., Lassila, O., and others, 2001. The semantic web. *Scientific american*, 284 (5), 28–37.

Bersini, H., Klatzmann, D., Six, A., and Thomas-Vaslin, V., 2012. State-Transition Diagrams for Biologists. *PLOS ONE*, 7 (7), e41165.

Bettembourg, C., Diot, C., and Dameron, O., 2014. Semantic Particularity Measure for Functional Characterization of Gene Sets Using Gene Ontology. *PLoS ONE*, 9 (1), e86525.

Bijalwan, Vishwanath, Vinay Kumar, Pinki Kumari, et Jordan Pascual. 2014. « KNN based machine learning approach for text and document mining ». *International Journal of Database Theory and Application* 7 (1): 61–70.

Bizer, C., Heath, T., and Berners-Lee, T., 2009. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 5 (3), 1–22.

Blanchet, Philippe. 2012. « La contextualisation entre sociolinguistique et sociodidactique : enjeux théoriques et méthodologiques ». *socles* 1 (1): 13-22.

Blaschke, C., Oliveros, J. C., and Valencia, A., 2001. Mining functional information associated with expression arrays. *Functional & Integrative Genomics*, 1 (4), 256–268.

Blei, David M., Andrew Y. Ng, et Michael I. Jordan. 2003. « Latent dirichlet allocation ». *Journal of machine Learning research* 3 (Jan): 993–1022.

Bodier, Stéphane, et Tiphaine Guerout. 2017. « Chapitre premier - Qu'est-ce que le web marketing en 2017 ? » *Que sais-je ?* 3e éd. (mars): 3-14.

Boole, G., 1854. *An Investigation of the Laws of Thought: On which are Founded the Mathematical Theories of Logic and Probabilities*. Dover Publications.

Borko, H., 1968. Information science: What is it? *In*: [online]. Presented at the American Documentation, 3. Available from: https://www.marilia.unesp.br/Home/Instituicao/Docentes/EdbertoFerneda/k---artigo-01.pdf [Accessed 26 Dec 2016].

Bornmann, L. and Mutz, R., 2015. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, 66 (11), 2215–2222.

Borst, P., 1997. Construction of engineering ontologies for knowledge sharing and reuse. Centre for Telematics and Information Technology, Enschede.

Boyd, D. and Crawford, K., 2012. Critical Questions for Big Data. *Information, Communication & Society*, 15 (5), 662–679.

Buermans, H. P. J. and den Dunnen, J. T., 2014. Next generation sequencing technology: Advances and applications. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, 1842 (10), 1932–1941.

Buyya, R., Yeo, C. S., Venugopal, S., Broberg, J., and Brandic, I., 2009. Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility. *Future Generation Computer Systems*, 25 (6), 599–616.

Calabrese, B. and Mario, C., 2015. *Bioinformatics and Microarray Data Analysis on the Cloud - Springer* [online]. Available from: http://link.springer.com.gate2.inist.fr/protocol/10.1007%2F7651_2015_236 [Accessed 22 Dec 2016].

Calon, A., Lonardo, E., Berenguer-Llergo, A., Espinet, E., Hernando-Momblona, X., Iglesias, M., Sevillano, M., Palomo-Ponce, S., Tauriello, D. V. F., Byrom, D., Cortina, C., Morral, C., Barceló, C., Tosi, S., Riera, A., Attolini, C. S.-O., Rossell, D., Sancho, E., and Batlle, E., 2015. Stromal gene expression defines poor-prognosis subtypes in colorectal cancer. *Nature Genetics*, 47 (4), 320–329.

Calvano, S. E., Xiao, W., Richards, D. R., Felciano, R. M., Baker, H. V., Cho, R. J., Chen, R. O., Brownstein, B. H., Cobb, J. P., Tschoeke, S. K., Miller-Graziano, C., Moldawer, L. L., Mindrinos, M. N., Davis, R. W., Tompkins, R. G., Lowry, S. F., and Program, I. and H. R. to I. L. S. C. R., 2005. A network-based analysis of systemic inflammation in humans. *Nature*, 437 (7061), 1032–1037.

Camefort, H., Génermont, J., L'héritier, P., and Lavergne, D., 2016. *REPRODUCTION, biologie* [online]. Encyclopædia Universalis. Available from: http://www.universalis.fr/encyclopedie/reproduction-biologie/ [Accessed 26 Dec 2016].

Cao, L., 2016. Data science: a comprehensive overview. *Submitted to ACM Computing Survey*, 1 (1), 1–37.

Cavalli-Sforza, L. L. and Feldman, M. W., 2003. The application of molecular genetic approaches to the study of human evolution. *Nature Genetics*, 33, 266–275.

Chandrasekaran, B., Josephson, J. R., Benjamins, V. R., and others, 1999. What are ontologies, and why do we need them? *IEEE Intelligent systems*, 14 (1), 20–26.

Charles F., H., 1960. The Origin Of Speech. *Scientific American*, September 1960, pp. 5–12.

Cheerla, N. and Gevaert, O., 2017. MicroRNA based Pan-Cancer Diagnosis and Treatment Recommendation. *BMC Bioinformatics*, 18 (1), 32.

Cherni, L., Fernandes, V., Pereira, J. B., Costa, M. D., Goios, A., Frigi, S., Yacoubi-Loueslati, B., Amor, M. B., Slama, A., Amorim, A., El Gaaied, A. B. A., and Pereira, L., 2009. Post-last glacial maximum expansion from Iberia to North Africa revealed by fine characterization of mtDNA H haplogroup in Tunisia. *American Journal of Physical Anthropology*, 139 (2), 253–260.

Cherni, L., Pakstis, A. J., Boussetta, S., Elkamel, S., Frigi, S., Khodjet-El-Khil, H., Barton, A., Haigh, E., Speed, W. C., Ben Ammar Elgaaied, A., Kidd, J. R., and Kidd, K. K., 2016. Genetic variation in Tunisia in the context of human diversity worldwide. *American Journal of Physical Anthropology*, 161 (1), 62–71.

Cherry, J. M., 2015. The Saccharomyces Genome Database: A Tool for Discovery. *Cold Spring Harbor Protocols*, 2015 (12), pdb.top083840.

Clark, J. W., 1901. *The care of books: an essay on the development of libraries and their fittings, from the earliest times to the end of the eighteenth century.* Cambridge: University press.

Cleveland, W. S., 2001. Data Science: an Action Plan for Expanding the Technical Areas of the Field of Statistics. *International Statistical Review*, 69 (1), 21–26.

Codd, E. F., 1970. A relational model of data for large shared data banks. *Communications of the ACM*, 13 (6), 377–387.

Collins, N. L., 2000. *The Library in Alexandria and the Bible in Greek* [online]. Brill. Available from: http://booksandjournals.brillonline.com/content/books/9789047400554 [Accessed 28 Dec 2016].

Constantin, L.-A. and Richter, N., 2006. *Bibliothéconomie: nouveau manuel complet pour l'arrangement, la construction et l'administration des bibliothèques*. Société d'histoire de la lecture.

De Courcy, R., 1992. Les systèmes d'information en réadaptation. *Québec, Réseau International CIDIH et facteurs environnementaux*, 1 (5), 7–10.

De Mauro, A., Greco, M., and Grimaldi, M., 2016. A formal definition of Big Data based on its essential features. *Library Review*, 65 (3), 122–135.

Dehak, N., Dehak, R., Kenny, P., Brümmer, N., Ouellet, P., and Dumouchel, P., 2009. Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification. *In*: *Interspeech* [online]. 1559–1562. Available from: http://www.crim.ca/perso/patrick.kenny/IS090079.PDF [Accessed 16 Jan 2017].

Dennis Jr, G., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C., Lempicki, R. A., and others, 2003. DAVID: database for annotation, visualization, and integrated discovery. *Genome biol*, 4 (5), P3.

Dérian, N., Bellier, B., Pham, H. P., Tsitoura, E., Kazazi, D., Huret, C., Mavromara, P., Klatzmann, D., and Six, A., 2016. Early Transcriptome Signatures from Immunized Mouse Dendritic Cells Predict Late Vaccine-Induced T-Cell Responses. *PLOS Computational Biology*, 12 (3), e1004801.

Dijkstra, E. W., 1971. *A Short Introduction to the Art of Programming*. Techn. Hogeschool.

Doherty, M. and Robertson, M. J., 2004. Some early Trends in Immunology. *TRENDS in Immunology*, 25 (12), 623–631.

Doms, A. and Schroeder, M., 2005. GoPubMed: exploring PubMed with the Gene Ontology. *Nucleic Acids Research*, 33 (Web Server issue), W783.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., and Picard, D., 1996. Density estimation by wavelet thresholding. *The Annals of Statistics*, 24 (2), 508–539.

Dragulanescu, N. G., 2003. De nouveaux modèles pour les sciences de l'information. *In*: [online]. Presented at the X° Colloque bilatéral franco-roumain, CIFSIC Université de Bucarest. Available from: https://halshs.archives-ouvertes.fr/file/index/docid/62201/filename/sic_00000514.pdf [Accessed 2 Jan 2017].

Driel, Marc A. van, Jorn Bruggeman, Gert Vriend, Han G. Brunner, et Jack A. M. Leunissen. 2006. « *A Text-Mining Analysis of the Human Phenome* ». *European Journal of Human Genetics* 14 (5): 535-42. doi:10.1038/sj.ejhg.5201585.

Dunham, I., Hunt, A. R., Collins, J. E., Bruskiewich, R., Beare, D. M., Clamp, M., Smink, L. J., Ainscough, R., Almeida, J. P., Babbage, A., Bagguley, C., Bailey, J., Barlow, K., Bates, K. N., Beasley, O., Bird, C. P., Blakey, S., Bridgeman, A. M., Buck, D., Burgess, J., Burrill, W. D., Burton, J., Carder, C., Carter, N. P., Chen, Y., Clark, G., Clegg, S. M., Cobley, V., Cole, C. G., Collier, R. E., Connor, R. E., Conroy, D., Corby, N., Coville, G. J., Cox, A. V., Davis, J., Dawson, E., Dhami, P. D., Dockree, C., Dodsworth, S. J., Durbin, R. M., Ellington, A., Evans, K. L., Fey, J. M., Fleming, K., French, L., Garner, A. A., Gilbert, J. G. R., Goward, M. E., Grafham, D., Griffiths, M. N., Hall, C., Hall, R., Hall-Tamlyn, G., Heathcott, R. W., Ho, S., Holmes, S., Hunt, S. E., Jones, M. C., Kershaw, J., Kimberley, A., King, A., Laird, G. K., Langford, C. F., Leversha, M. A., Lloyd, C., Lloyd, D. M., Martyn, I. D., Mashreghi-Mohammadi, M., Matthews, L., McCann, O. T., McClay, J., McLaren, S., McMurray, A. A., Milne, S. A., Mortimore, B. J., Odell, C. N., Pavitt, R., Pearce, A. V., Pearson, D., Phillimore, B. J., Phillips, S. H., Plumb, R. W., Ramsay, H., Ramsey, Y., Rogers, L., Ross, M. T., Scott, C. E., Sehra, H. K., Skuce, C. D., Smalley, S., Smith, M. L., Soderlund, C., Spragon, L., Steward, C. A., Sulston, J. E., Swann, R. M., Vaudin, M., Wall, M., Wallis, J. M., Whiteley, M. N., Willey, D., Williams, L., Williams, S., Williamson, H., Wilmer, T. E., Wilming, L., Wright, C. L., Hubbard, T., Bentley, D. R., Beck, S., Rogers, J., Shimizu, N., Minoshima, S., Kawasaki, K., Sasaki, T., Asakawa, S., Kudoh, J., Shintani, A., Shibuya, K., Yoshizaki, Y., Aoki, N., Mitsuyama, S., Roe, B. A., Chen, F., Chu, L., Crabtree, J., Deschamps, S., Do, A., Do, T., Dorman, A., Fang, F., Fu, Y., Hu, P., Hua, A., Kenton, S., Lai, H., Lao, H. I., Lewis, J., Lewis, S., Lin, S.-P., Loh, P., Malaj, E., Nguyen, T., Pan, H., Phan, S., Qi, S., Qian, Y., Ray, L., Ren, Q., Shaull, S., Sloan, D., Song, L., Wang, Q., Wang, Y., Wang, Z., White, J., Willingham, D., Wu, H., Yao, Z., Zhan, M., Zhang, G., Chissoe, S., Murray, J., Miller, N., Minx, P., Fulton, R., Johnson, D., Bemis, G., Bentley, D., Bradshaw, H., Bourne, S., Cordes, M., Du, Z., Fulton, L., Goela, D.,

Graves, T., Hawkins, J., Hinds, K., Kemp, K., Latreille, P., Layman, D., Ozersky, P., Rohlfing, T., Scheet, P., Walker, C., Wamsley, A., Wohldmann, P., Pepin, K., Nelson, J., Korf, I., Bedell, J. A., Hillier, L., Mardis, E., Waterston, R., Wilson, R., Emanuel, B. S., Shaikh, T., Kurahashi, H., Saitta, S., Budarf, M. L., McDermid, H. E., Johnson, A., Wong, A. C. C., Morrow, B. E., Edelmann, L., Kim, U. J., Shizuya, H., Simon, M. I., Dumanski, J. P., Peyrard, M., Kedra, D., Seroussi, E., Fransson, I., Tapia, I., Bruder, C. E., and O'Brien, K. P., 1999. The DNA sequence of human chromosome 22. *Nature*, 402 (6761), 489–495.

Dunsch, S., 2016. Okun's Law and Youth Unemployment in Germany and Poland. *International Journal of Management and Economics* [online], 49 (1). Available from: http://www.degruyter.com/view/j/ijme.2016.49.issue-1/ijme-2016-0003/ijme-2016-0003.xml [Accessed 8 Jan 2017].

Dutkowski, J., Ono, K., Kramer, M., Yu, M., Pratt, D., Demchak, B., and Ideker, T., 2014. NeXO Web: the NeXO ontology database and visualization platform. *Nucleic Acids Research*, 42 (D1), D1269–D1274.

Eberly, L. E., 2007. Multiple Linear Regression. *In*: Ambrosius, W. T., ed. *Topics in Biostatistics* [online]. Totowa, NJ: Humana Press, 165–187. Available from: http://dx.doi.org/10.1007/978-1-59745-530-5_9.

Elshamy, H., 2013. The Relationship Between Unemployment and Output in Egypt. *Procedia - Social and Behavioral Sciences*, 81, 22–26.

Eppig, J. T., Smith, C. L., Blake, J. A., Ringwald, M., Kadin, J. A., Richardson, J. E., and Bult, C. J., 2017. Mouse Genome Informatics (MGI): Resources for Mining Mouse Genetic, Genomic, and Biological Data in Support of Primary and Translational Research. *Methods in Molecular Biology (Clifton, N.J.)*, 1488, 47–73.

Ermann, J., Rao, D. A., Teslovich, N. C., Brenner, M. B., and Raychaudhuri, S., 2015. Immune cell profiling to guide therapeutic decisions in rheumatic diseases. *Nature Reviews Rheumatology*, 11 (9), 541–551.

Falk, D., 2004. Prelinguistic evolution in early hominins: Whence motherese? *Behavioral and Brain Sciences*, 27 (04), 491–503.

Fang, H., Yamaguchi, R., Liu, X., Daigo, Y., Yew, P. Y., Tanikawa, C., Matsuda, K., Imoto, S., Miyano, S., and Nakamura, Y., 2014. Quantitative T cell repertoire analysis by deep cDNA sequencing of T cell receptor α and β chains using next-generation sequencing (NGS). *OncoImmunology*, 3 (12), e968467.

Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. 1996. « The KDD Process for Extracting Useful Knowledge from Volumes of Data ». Commun. ACM 39 (11): 27–34. doi:10.1145/240455.240464.

Fayyad, U. and Stolorz, P., 1997. Fouille des données and KDD: Promise and challenges. *Future Generation Computer Systems*, 13 (2), 99–115.

Feldman, R. and Sanger, J., 2007. *The fouille des textes Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press.

Fisher, R. A., 1918. The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Transactions of the Royal Society of Edinburgh*, 52, 399–433.

Flajnik, M. F. and Kasahara, M., 2010. Origin and evolution of the adaptive immune system: genetic events and selective pressures. *Nature Reviews Genetics*, 11 (1), 47–59.

Floridi, L., 2005. Semantic Conceptions of Information. [online]. Available from: https://seop.illc.uva.nl/entries/information-semantic/ [Accessed 4 Mar 2017].

Fodor, S. P., Read, J. L., Pirrung, M. C., Stryer, L., Lu, A. T., and Solas, D., 1991. Light-directed, spatially addressable parallel chemical synthesis. *Science*, 251 (4995), 767–773.

Frisch, Matthias, Bernward Klocke, Manuela Haltmeier, et Kornelie Frech. 2009. « LitInspector: literature and signal transduction pathway mining in PubMed abstracts ». *Nucleic Acids Research* 37 (suppl_2): W135-40. doi:10.1093/nar/gkp303.

Galton, F., 1886. Regression Towards Mediocrity in Hereditary Stature. *Anthropological Miscellanea*, 15, 246–263.

Gan, M., Dou, X., and Jiang, R., 2013. From ontology to semantic similarity: calculation of ontology-based semantic similarity. *TheScientificWorldJournal*, 2013, 793091.

Gattinoni, L., Speiser, D. E., Lichterfeld, M., and Bonini, C., 2017. T memory stem cells in health and disease. *Nature Medicine* [online], 23. Available from: http://www.readcube.com/articles/10.1038/nm.4241 [Accessed 14 Jan 2017].

Getoor, Lise, et Ben Taskar. 2007. *Introduction to Statistical Relational Learning*. MIT Press.

Gillings, Michael R., Martin Hilbert, et Darrell J. Kemp. 2016. « Information in the Biosphere: Biological and Digital Worlds ». *Trends in Ecology & Evolution* 31 (3): 180-89. doi:10.1016/j.tree.2015.12.013.

Gilmont, J.-F., 2004. *Une introduction à l'histoire du livre et de la lecture : du manuscrit à l'ère électronique*. Editions du CEFAL.

Glowniak, J., 1998. History, structure, and function of the internet. *Seminars in Nuclear Medicine*, 28 (2), 135–144.

Goldberg, D. E., 1989. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley Publishing Company.

Goldstine, H. H. and Goldstine, A., 1946. The Electronic Numerical Integrator and Computer (ENIAC). *Mathematical Tables and Other Aids to Computation*, 2 (15), 97–110.

Gonzalez, G. H., Tahsin, T., Goodale, B. C., Greene, A. C., and Greene, C. S., 2016. Recent Advances and Emerging Applications in Text and Fouille des données for Biomedical Discovery. *Briefings in Bioinformatics*, 17 (1), 33–42.

Grace, C. and Nacheva, E. P., 2012. Significance Analysis of Microarrays (SAM) Offers Clues to Differences Between the Genomes of Adult Philadelphia Positive ALL and the Lymphoid Blast Transformation of CML. *Cancer Informatics*, 11, 173–183.

Green, P., 1993. *Alexander to Actium: The Historical Evolution of the Hellenistic Age*. University of California Press.

Green, R. E., Krause, J., Ptak, S. E., Briggs, A. W., Ronan, M. T., Simons, J. F., Du, L., Egholm, M., Rothberg, J. M., Paunovic, M., and Pääbo, S., 2006. Analysis of one million base pairs of Neanderthal DNA. *Nature*, 444 (7117), 330–336.

Green, S., 2016. *Philosophy of Systems Biology: Perspectives from Scientists and Philosophers*. Springer.

Grinberg-Bleyer, Y., Saadoun, D., Baeyens, A., Billiard, F., Goldstein, J. D., Grégoire, S., Martin, G. H., Elhage, R., Derian, N., Carpentier, W., Marodon, G., Klatzmann, D.,

Piaggio, E., and Salomon, B. L., 2010. Pathogenic T cells have a paradoxical protective effect in murine autoimmune diabetes by boosting Tregs. *The Journal of Clinical Investigation*, 120 (12), 4558–4568.

Groza, T., Köhler, S., Doelken, S., Collier, N., Oellrich, A., Smedley, D., Couto, F. M., Baynam, G., Zankl, A., and Robinson, P. N., 2015. Automatic concept recognition using the Human Phenotype Ontology reference and test suite corpora. *Database: The Journal of Biological Databases and Curation* [online], 2015. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4343077/ [Accessed 18 Aug 2015].

Gruber, T. R., 1993. A translation approach to portable ontology specifications. *Knowledge acquisition*, 5 (2), 199–220.

Gruber, T. R., 1995. Toward principles for the design of ontologies used for knowledge sharing? *International Journal of Human-Computer Studies*, 43 (5), 907–928.

Guarino, N., Oberle, D., and Staab, S., 2009. What Is an Ontology ? *In*: Staab, S. and Studer, R., eds. *Handbook on Ontologies* [online]. Berlin, Heidelberg: Springer Berlin Heidelberg, 1–17. Available from: http://link.springer.com/10.1007/978-3-540-92673-3 [Accessed 3 Jan 2017].

Guerrero, Manuel, Francisco G. Montoya, Raúl Baños, Alfredo Alcayde, et Consolación Gil. 2017. « Adaptive community detection in complex networks using genetic algorithms ».*Neurocomputing* 266 (novembre): 101-13. doi:10.1016/j.neucom.2017.05.029.

Guo, X., Shriver, C. D., Hu, H., and Liebman, M. N., 2005. Analysis of Metabolic and Regulatory Pathways through Gene Ontology-Derived Semantic Similarity Measures. *AMIA Annual Symposium Proceedings*, 2005, 972.

Guo, X., Liu, R., Shriver, C. D., Hu, H., and Liebman, M. N., 2006. Assessing semantic similarity measures for the characterization of human regulatory pathways. *Bioinformatics*, 22 (8), 967–973.

Hajjej, A., Almawi, W. Y., Hattab, L., El-Gaaied, A., and Hmida, S., 2016. The investigation of the origin of Southern Tunisians using HLA genes. *Journal of Human Genetics* [online]. Available from: http://www.nature.com.gate2.inist.fr/jhg/journal/vaop/ncurrent/full/jhg2016146a.html [Accessed 26 Dec 2016].

Han, J., Kamber, M., and Pei, J., 2012. 1 - Introduction. *In*: *Fouille des données (Third Edition)* [online]. Boston: Morgan Kaufmann, 1–38. Available from: http://www.sciencedirect.com/science/article/pii/B9780123814791000010 [Accessed 5 Jan 2017].

Hanusch, M., 2012. *Jobless growth ? Okun's law in East Asia* [online]. The World Bank. No. WPS6156. Available from: http://documents.banquemondiale.org/curated/fr/683701468036885817/Jobless-growth-Okuns-law-in-East-Asia [Accessed 8 Jan 2017].

Hartigan, J. A. and Wong, M. A., 1979. A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28 (1), 100–108.

Hassani, H., Huang, X., Silva, E. S., and Ghodsi, M., 2016. A Review of Fouille des données Applications in Crime. *ResearchGate* [online], 9 (3). Available from: https://www.researchgate.net/publication/301579904_A_Review_of_Data_Mining_Applications_in_Crime [Accessed 8 Jan 2017].

Hastie, T., Tibshirani, R., and Friedman, J., 2013. *Elements of Statistical Learning: fouille des données, inference, and prediction. 2nd Edition.* [online]. 2nd ed. Springer Science & Business Media. Available from: http://statweb.stanford.edu/~tibs/ElemStatLearn/ [Accessed 12 Jan 2017].

Hasty, P., Campisi, J., Hoeijmakers, J., Steeg, H. van, and Vijg, J., 2003. Aging and Genome Maintenance: Lessons from the Mouse? *Science*, 299 (5611), 1355–1359.

Hedrich, R., Salvador-Recatalà, V., and Dreyer, I., 2016. Electrical Wiring and Long-Distance Plant Communication. *Trends in Plant Science*, 21 (5), 376–387.

Heidorn, P. B., Palmer, C. L., and Wright, D., 2007. Biological information specialists for biological informatics. *Journal of Biomedical Discovery and Collaboration*, 2, 1.

Henderson, J. V., Storeygard, A., and Weil, D. N., 2012. MEASURING ECONOMIC GROWTH FROM OUTER SPACE. *The American economic review*, 102 (2), 994.

Hershey, A. D. and Chase, M., 1952. Independent Functions of Viral Protein and Nucleic Acid in Growth of Bacteriophage. *The Journal of General Physiology*, 36 (1), 39–56.

Hilbert, M. and López, P., 2011. The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332 (6025), 60–65.

Holland, J. H., 1992. *Adaptation in Natural and Artificial Systems* [online]. MIT Press. Available from: https://mitpress.mit.edu/books/adaptation-natural-and-artificial-systems [Accessed 19 Dec 2016].

Holmes, R. K. and Jobling, M. G., 1996. Genetics. *In*: Baron, S., ed. *Medical Microbiology* [online]. Galveston (TX): University of Texas Medical Branch at Galveston. Available from: http://www.ncbi.nlm.nih.gov/books/NBK7908/ [Accessed 26 Dec 2016].

Hou, Q., De Geest, P., Vranken, W. F., Heringa, J., and Feenstra, K. A., 2017. Seeing the Trees through the Forest: Sequence-based Homo- and Heteromeric Protein-protein Interaction sites prediction using Random Forest. *Bioinformatics (Oxford, England)*.

Huang, S. and Fidrmuc, J., 2016. Unemployment and the Speed of Transition in China. *Asian Economic Papers*, 15 (1), 156–170.

Humphreys, B. L., Lindberg, D. A. B., Schoolman, H. M., and Barnett, G. O., 1998. The Unified Medical Language System: An Informatics Research Collaboration. *Journal of the American Medical Informatics Association : JAMIA*, 5 (1), 1.

Information, N. C. for B., Pike, U. S. N. L. of M. 8600 R., MD, B., and Usa, 20894, 1999. The FlyBase database of the Drosophila Genome Projects and community literature. The FlyBase Consortium. *Nucleic Acids Research*, 27 (1), 85.

Iniesta, R., Stahl, D., and McGuffin, P., 2016. Machine learning, statistical learning and the future of biological research in psychiatry. *Psychological Medicine*, 46 (12), 2455–2465.

Ivanović, M. and Budimac, Z., 2014. An overview of ontologies and data resources in medical domains. *Expert Systems with Applications*, 41 (11), 5158–5166.

Jacob, F. and Monod, J., 1961. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology*, 3 (3), 318–356.

Jaffe, Andrew E., Ran Tao, Alexis L. Norris, Marc Kealhofer, Abhinav Nellore, Joo Heon Shin, Dewey Kim, et al. 2017. « QSVA Framework for RNA Quality Correction in

Differential Expression Analysis ». *Proceedings of the National Academy of Sciences* 114(27): 7130-35. doi:10.1073/pnas.1617384114.

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., and Ermon, S., 2016. Combining satellite imagery and machine learning to predict poverty. *Science*, 353 (6301), 790–794.

Jelier, R., Schuemie, M. J., Veldhoven, A., Dorssers, L. C., Jenster, G., and Kors, J. A., 2008. ANNI 2.0: a multipurpose text-mining tool for the life sciences. *Genome Biology*, 9, R96.

Jensen, L. J., Jensen, T. S., de Lichtenberg, U., Brunak, S., and Bork, P., 2006. Co-evolution of transcriptional and post-translational cell-cycle regulation. *Nature*, 443 (7111), 594–597.

Jeong, J. H., Resop, J. P., Mueller, N. D., Fleisher, D. H., Yun, K., Butler, E. E., Timlin, D. J., Shim, K.-M., Gerber, J. S., Reddy, V. R., and Kim, S.-H., 2016. Random Forests for Global and Regional Crop Yield Predictions. *PLOS ONE*, 11 (6), e0156571.

Jessulat, M., Pitre, S., Gui, Y., Hooshyar, M., Omidi, K., Samanfar, B., Tan, L. H., Alamgir, M., Green, J., Dehne, F., and Golshani, A., 2011. Recent advances in protein–protein interaction prediction: experimental and computational methods. *Expert Opinion on Drug Discovery*, 6 (9), 921–935.

Jim, L., 2000. The Robustness of Okun's Law: Evidence from OECD countries. *Journal of Macroeconomic*, 22 (2), 331–356.

Johnson, S. C., 1967. Hierarchical clustering schemes. *Psychometrika*, 32 (3), 241–254.

Jonquet, C., Shah, N., Youn, C., Callendar, C., Storey, M.-A., and Musen, M., 2009. NCBO annotator: semantic annotation of biomedical data. *In*: *International Semantic Web Conference* [online]. Available from: http://www.lirmm.fr/~jonquet/publications/documents/Demo-ISWC09-Jonquet.pdf [Accessed 6 Nov 2015].

Kalendar, R., Belyayev, A., Zachepilo, T., Vaido, A., Maidanyuk, D., Schulman, A. H., and Dyuzhikova, N., 2017. Copy-number variation of housekeeping gene rpl13a in rat strains selected for nervous system excitability. *Molecular and Cellular Probes*, 1–5.

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K., 2017. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, 45 (D1), D353–D361.

Kanehisa, M. and Goto, S., 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28 (1), 27–30.

Kawai, J., Shinagawa, A., Shibata, K., Yoshino, M., Itoh, M., Ishii, Y., Arakawa, T., Hara, A., Fukunishi, Y., Konno, H., Adachi, J., Fukuda, S., Aizawa, K., Izawa, M., Nishi, K., Kiyosawa, H., Kondo, S., Yamanaka, I., and Saito, T., 2001. Functional annotation of a full-length mouse cDNA collection. *Nature*, 409 (6821), 685–690.

Khorana, H. G., 1968. Nucleic acid synthesis in the study of the genetic code. *Nobel Lectures: Physiology or Medicine (1963–1970)*, 341–369.

Kibbe, W. A., Arze, C., Felix, V., Mitraka, E., Bolton, E., Fu, G., Mungall, C. J., Binder, J. X., Malone, J., Vasant, D., Parkinson, H., and Schriml, L. M., 2014. Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic Acids Research*.

Kishore Kumar, R., Poonkuzhali, G., and Sudhakar, P., 2012. Comparative Study on Email Spam Classifier using Fouille des données Techniques. *In*: . Presented at the Proceedings of the International MultiConference of Engineers and Computer Scientists: IMECS 2012, Hong Kong.

Kitano, H., 2002. Computational systems biology. *Nature*, 420 (6912), 206–210.

Kowalski, R., 1974. Predicate logic as programming language. *Information processing*, (74), 569–574.

Krallinger, M., Erhardt, R. A.-A., and Valencia, A., 2005. Text-mining approaches in molecular biology and biomedicine. *Drug Discovery Today*, 10 (6), 439–445.

Krämer, A., Green, J., Pollard, J., and Tugendreich, S., 2014. Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinformatics*, 30 (4), 523–530.

Kreil, D. P. and Russell, R. R., 2005. There is no silver bullet--a guide to low-level data transforms and normalisation methods for microarray data. *Briefings in Bioinformatics*, 6 (1), 86–97.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.-F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., Bastide, M. de la, Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H.-C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G. R., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F. A., Stupka, E., Szustakowki, J., Thierry-Mieg, D.,

Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S.-P., Yeh, R.-F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Patrinos, A., and Morgan, M. J., 2001. Initial sequencing and analysis of the human genome. *Nature*, 409 (6822), 860–921.

Larose, D. T., 2005. *Discovering knowledge in data: an introduction to fouille des données*. Hoboken, N.J: Wiley-Interscience.

Leleu-Merviel, S. and Useille, P., 2008. *Quelques révisions du concept d'information* [online]. Lavoisier. Available from: https://hal.archives-ouvertes.fr/hal-00695777/document [Accessed 4 Mar 2017].

Lemma, S., Avnet, S., Salerno, M., Chano, T., and Baldini, N., 2016. Identification and Validation of Housekeeping Genes for Gene Expression Analysis of Cancer Stem Cells. *PLOS ONE*, 11 (2), e0149481.

Lena, P. D., Domeniconi, G., Margara, L., and Moro, G., 2015. GOTA: GO term annotation of biomedical literature. *BMC Bioinformatics* [online], 16. Available from: https://www-ncbi-nlm-nih-gov.gate2.inist.fr/pmc/articles/PMC4625458/ [Accessed 24 Jan 2017].

Lenoir, T. and Giannella, E., 2006. The emergence and diffusion of DNA microarray technology. *Journal of Biomedical Discovery and Collaboration*, 1, 11.

Li, A., Yu, J., Kim, H., Wolfgang, C. L., Canto, M. I., Hruban, R. H., and Goggins, M., 2013. MicroRNA array analysis finds elevated serum miR-1290 accurately distinguishes patients with low-stage pancreatic cancer from healthy and disease controls. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 19 (13), 3600.

Li, J., Klughammer, J., Farlik, M., Penz, T., Spittler, A., Barbieux, C., Berishvili, E., Bock, C., and Kubicek, S., 2016. Single- cell transcriptomes reveal characteristic features of human pancreatic islet cell types. *EMBO reports*, 17 (2), 178–187.

Li, Z. and Liu, J., 2016. A multi-agent genetic algorithm for community detection in complex networks. *Physica A: Statistical Mechanics and its Applications*, 449, 336–347.

Lin, D., 1998. An Information-Theoretic Definition of Similarity. *In*: *Proceedings of the Fifteenth International Conference on Machine Learning* [online]. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 296–304. Available from: http://dl.acm.org/citation.cfm?id=645527.657297 [Accessed 19 Dec 2016].

Litman, G. W., Cannon, J. P., and Dishaw, L. J., 2005. RECONSTRUCTING IMMUNE PHYLOGENY: NEW PERSPECTIVES. *Nature reviews. Immunology*, 5 (11), 866.

Liu, X., Yu, X., Zack, D. J., Zhu, H., and Qian, J., 2008. TiGER: A database for tissue-specific gene expression and regulation. *BMC Bioinformatics*, 9 (1), 271.

Longo, G., Miquel, P.-A., Sonnenschein, C., and Soto, A. M., 2012. Is information a proper observable for biological organization? *Progress in Biophysics and Molecular Biology*, 109 (3), 108–114.

Love, B. C., 2002. Comparing supervised and unsupervised category learning. *Psychonomic Bulletin & Review*, 9 (4), 829–835.

Lu, Z., 2011. PubMed and beyond: a survey of web tools for searching biomedical literature. *Database*, 2011, baq036.

Marill, T. and Roberts, L. G., 1966. Toward a cooperative network of time-shared computers. *In*: *Proceedings of the November 7-10, 1966, fall joint computer conference* [online]. ACM, 425–431. Available from: http://dl.acm.org/citation.cfm?id=1464336 [Accessed 30 Dec 2016].

Masson, J., 2006. Apollo 11 Cave in Southwest Namibia: Some Observations on the Site and Its Rock Art. *The South African Archaeological Bulletin*, 61 (183), 76–89.

Masys, D. R., 2001. Linking microarray data to the literature. *Nature Genetics*, 28 (1), 9–10.

Maxam, A. M. and Gilbert, W., 1977. A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 74 (2), 560.

Mazandu, Gaston K., et Nicola J. Mulder. 2014. « Information Content-Based Gene Ontology Functional Similarity Measures: Which One to Use for a Given Biological Data Type? » *PLOS ONE* 9 (12): e113859. doi:10.1371/journal.pone.0113859.

McLachlan, G. J., Bean, R. W., and Ng, S. K., 2017. Clustering. *In*: Keith, J. M., ed. *Bioinformatics* [online]. New York, NY: Springer New York, 345–362. Available from: http://link.springer.com/10.1007/978-1-4939-6613-4_19 [Accessed 22 Dec 2016].

Meldrum, D., 2000a. Automation for Genomics, Part One: Preparation for Sequencing. *Genome Research*, 10 (8), 1081–1092.

Meldrum, D., 2000b. Automation for Genomics, Part Two: Sequencers, Microarrays, and Future Trends. *Genome Research*, 10 (9), 1288–1303.

Mendel, J. G., 1865. Recherches sur des hybrides végétaux. *In*: Chappelier, A., tran. [online]. Presented at the Verhandlungen des naturforschenden Vereines in Brünn, Brünn, Allemagne, 3–47. Available from: https://fr.wikisource.org/wiki/Recherches_sur_des_hybrides_v%C3%A9g%C3%A9taux [Accessed 13 Jan 2017].

Mitchell, T. M., 2006. *The discipline of machine learning* [online]. Carnegie Mellon University, School of Computer Science, Machine Learning Department. Available from: http://www-cgi.cs.cmu.edu/~tom/pubs/MachineLearningTR.pdf [Accessed 8 Jan 2017].

Mohri, M., Rostamizadeh, A., and Talwalkar, A., 2012. *Foundations of Machine Learning*. MIT Press.

Molla, Y. B., Rawlins, B., Makanga, P. T., Cunningham, M., Ávila, J. E. H., Ruktanonchai, C. W., Singh, K., Alford, S., Thompson, M., Dwivedi, V., Moran, A. C., and Matthews, Z., 2017. Geographic information system for improving maternal and newborn health: recommendations for policy and programs. *BMC pregnancy and childbirth*, 17 (1), 26.

Morgan, T. H., 1910. Chromosomes and Heredity. *The American Naturalist*, 44 (524), 449–496.

Mungall, C. J., Torniai, C., Gkoutos, G. V., Lewis, S. E., and Haendel, M. A., 2012. Uberon, an integrative multi-species anatomy ontology. *Genome Biology*, 13, R5.

Muraro, Daniele, et Alison Simmons. 2016. « An integrative analysis of gene expression and molecular interaction data to identify dys-regulated sub-networks in inflammatory bowel disease ».*BMC Bioinformatics* 17 (janvier): 42. doi:10.1186/s12859-016-0886-z.

Navarro, E. and Cazabet, R., 2010. Détection de communautés, étude comparative sur

graphes réels. In: [online]. Presented at the *Journées Modèles et l'Analyse des Réseaux : Approches Mathématiques et Informatique*, Toulouse. Available from: https://enavarro.me/pages/publications.html [Accessed 18 May 2017].

Needham, J. and Tsuen-Hsuin, T., 1985. *Science and Civilisation in China: Volume 5, Chemistry and Chemical Technology, Part 1, Paper and Printing*. Cambridge University Press.

Nehar-Belaid, D., Courau, T., Dérian, N., Florez, L., Ruocco, M. G., and Klatzmann, D., 2016. Regulatory T Cells Orchestrate Similar Immune Evasion of Fetuses and Tumors in Mice. *The Journal of Immunology*, 196 (2), 678–690.

Nhavoto, José António, et Åke Grönlund. 2014. « Mobile Technologies and Geographic Information Systems to Improve Health Care Systems: A Literature Review ». JMIR MHealth and UHealth 2 (2). doi:10.2196/mhealth.3216.

Nilsson, N. J., 1991. Logic and artificial intelligence. *Artificial intelligence*, 47 (1–3), 31–56.

Nirenberg, M. and Leder, P., 1964. RNA Codewords and Protein Synthesis. *Science*, 145 (3639), 1399–1407.

Nishimura, M., Naito, S., and Yokoi, T., 2004. Tissue-specific mRNA Expression Profiles of Human Nuclear Receptor Subfamilies. *Drug Metabolism and Pharmacokinetics*, 19 (2), 135–149.

Nikos, Smyrnaios. 2016. « L'effet GAFAM : stratégies et logiques de l'oligopole de l'internet. » *Communication & langages* 2016 (188): 61-83. doi:10.4074/S0336150016012047.

Noy, N. F., Shah, N. H., Whetzel, P. L., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Rubin, D. L., Storey, M.-A., Chute, C. G., and Musen, M. A., 2009. BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Research*, 37 (Web Server issue), W170-173.

Okun, A. M., 1962. Potential GNP: it's measurement and significance. *Proceedings of the Business and Economics Section*, 98–103.

Ola-David, O., Oluwatobi, S., and Ogundipe, A., 2016. Output and Unemployment Relationship: How Applicable Is the Okun's Law to Nigeria? *ResearchGate*, 11 (8), 1422–1427.

Oliveros, J. C., Blaschke, C., Herrero, J., Dopazo, J., and Valencia, A., 2000. Expression profiles and biological function. *Genome Informatics. Workshop on Genome Informatics*, 11, 106–117.

Orgogozo, V., Morizot, B., and Martin, A., 2015. The differential view of genotype–phenotype relationships. *Frontiers in Genetics* [online], 6. Available from: http://journal.frontiersin.org/article/10.3389/fgene.2015.00179/abstract [Accessed 27 Feb 2017].

Oriol, J.-C., 2007. Formation à la statistique par la pratique d'enquêtes par questionnaires et la simulation : étude didactique d'une expérience d'enseignement dans un département d'IUT. phdthesis. [online]. Université Lumière - Lyon II. Available from: https://tel.archives-ouvertes.fr/tel-00191166/document [Accessed 4 Jan 2017].

Otlet, P., 1934. *Traité de documentation : le livre sur le livre (théorie et pratique)* [online]. Edicioncs Mundancum, Palais Mondial. D. Van Keerberghen & fils. Available from:

http://lib.ugent.be/fulltxt/handle/1854/5612/Traite_de_documentation_ocr.pdf [Accessed 29 Dec 2016].

Palmer, C., Diehn, M., Alizadeh, A. A., and Brown, P. O., 2006. Cell-type specific gene expression profiles of leukocytes in human peripheral blood. *BMC Genomics*, 7 (1), 115.

Pan, S. J. and Yang, Q., 2010. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22 (10), 1345–1359.

Parrini-Alemanno, S., 2007. Contexte et contextualisation dans l'approche qualitative de la communication organisationnelle. *Recherche Qualitative*, (3), 335–349.

Partyko, Z.V. 2009. « The Modern Paradigm of Information Science: Informology ». *Automatic Documentation and Mathematical Linguistics* 43 (6): 311-20. doi:10.3103/S0005105509060016.

Parzen, Emanuel. 1962. « On Estimation of a Probability Density Function and Mode ». *The Annals of Mathematical Statistics* 33 (3): 1065-76. doi:10.1214/aoms/1177704472.

Paul, A. K., Shill, P. C., and Kundu, A., 2016. A multi-objective genetic algorithm based fuzzy relational clustering for automatic microarray cancer data clustering. *In*: *2016 5th International Conference on Informatics, Electronics and Vision (ICIEV)*. Presented at the 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV), 454–459.

Pekar, V. and Staab, S., 2002. Taxonomy Learning: Factoring the Structure of a Taxonomy into a Semantic Classification Decision. *In*: *Proceedings of the 19th International Conference on Computational Linguistics - Volume 1* [online]. Stroudsburg, PA, USA: Association for Computational Linguistics, 1–7. Available from: http://dx.doi.org/10.3115/1072228.1072318 [Accessed 22 Nov 2016].

Peregrín-Alvarez, J. M., Sanford, C., and Parkinson, J., 2009. The conservation and evolutionary modularity of metabolism. *Genome Biology*, 10 (6), R63.

Perreault, C. and Mathew, S., 2012. Dating the Origin of Language Using Phonemic Diversity. *PLOS ONE*, 7 (4), e35289.

Pesquita, C., Faria, D., Falcão, A. O., Lord, P., and Couto, F. M., 2009. Semantic Similarity in Biomedical Ontologies. *PLOS Computational Biology*, 5 (7), e1000443.

Pham, H. P., 2013. *L'importance de la variabilité inter-individuelle dans l'étude de la dynamique et de la diversité du répertoire des lymphocytes T au cours du vieillessement* [online]. Paris 6. Available from: http://www.theses.fr/2013PA066156 [Accessed 30 Oct 2015].

Pham, H.-P., Dérian, N., Chaara, W., Bellier, B., Klatzmann, D., and Six, A., 2014. A novel strategy for molecular signature discovery based on independent component analysis. *International Journal of Fouille des données and Bioinformatics*, 9 (3), 277–304.

Pisanski, K., Cartei, V., McGettigan, C., Raine, J., and Reby, D., 2016. Voice Modulation: A Window into the Origins of Human Vocal Control? *Trends in Cognitive Sciences*, 20 (4), 304–318.

Pizzuti, C., 2008. GA-Net: A Genetic Algorithm for Community Detection in Social Networks. *In*: Rudolph, G., Jansen, T., Beume, N., Lucas, S., and Poloni, C., eds. *Parallel Problem Solving from Nature – PPSN X* [online]. Presented at the International Conference on Parallel Problem Solving from Nature, Springer Berlin Heidelberg, 1081–1090.

Available from: http://link.springer.com.gate2.inist.fr/chapter/10.1007/978-3-540-87700-4_107 [Accessed 17 Dec 2016].

Prachowny, M. F. J., 1993. Okun's Law: Theoretical Foundations and Revised Estimates. *ResearchGate*, 75 (2), 331–36.

Provost, F. and Fawcett, T., 2013. Data Science and its Relationship to Big Data and Data-Driven Decision Making. *Big Data*, 1 (1), 51–59.

Quackenbush, J., 2002. Microarray data normalization and transformation. *Nature Genetics*, 32, 496–501.

Rafael, C. and Hjørland, B., 2003. Concept of Information. *Annual Review of Information Science and Technology*, 37 (8), 343–411.

Rayward, W. B., 1997. The Origins of Information Science and the International Institute of Bibliography/ International Federation for Information and Documentation(FID). *Journal of the American Society for Information Science*, 48, 289–300.

Reed, D. L., Smith, V. S., Hammond, S. L., Rogers, A. R., and Clayton, D. H., 2004. Genetic Analysis of Lice Supports Direct Contact between Modern and Archaic Humans. *PLOS Biology*, 2 (11), e340.

Resnik, P., 1995. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. *In*: *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1* [online]. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 448–453. Available from: http://dl.acm.org/citation.cfm?id=1625855.1625914 [Accessed 19 Dec 2016].

Richard D., T., 1992. The Structure and Function Of the Immune System And Mechanisms of Immunotoxicity. *In*: Kate, K., ed. *Biologic Markers in Immunotoxicology* [online]. Washington, D.C. USA: National Academies Press (US), 23–32. Available from: https://www-ncbi-nlm-nih-gov.gate2.inist.fr/books/NBK235674/ [Accessed 14 Jan 2017].

Rincy, J. and Varghese, S. C., 2015. Prediction of election result by enhanced sentiment analysis on twitter data using classifier ensemble Approach. *In*: [online]. Presented at the Control Communication & Computing India, Trivandrum, India: IEEE, 6386641. Available from: http://ieeexplore.ieee.org/abstract/document/7684133/ [Accessed 13 Jan 2017].

Riz, I., Hawley, T. S., and Hawley, R. G., 2015. KLF4-SQSTM1/p62-associated prosurvival autophagy contributes to carfilzomib resistance in multiple myeloma models. *Oncotarget*, 6 (17), 14814.

Robert G., B., 2003. THE EARLIEST EVIDENCE OF PALAEOART. *Rock Art Research: The Journal of the Australian Rock Art Research Association (AURA)*, 20 (2), 3–28.

Rosario, B. and Hearst, M. A., 2004. Classifying semantic relations in bioscience texts. *In*: *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics* [online]. Association for Computational Linguistics, 430. Available from: http://dl.acm.org/citation.cfm?id=1219010 [Accessed 30 Oct 2015].

Rosenzwajg, M., Churlaud, G., Mallone, R., Six, A., Dérian, N., Chaara, W., Lorenzon, R., Long, S. A., Buckner, J. H., Afonso, G., Pham, H.-P., Hartemann, A., Yu, A., Pugliese, A., Malek, T. R., and Klatzmann, D., 2015. Low-dose interleukin-2 fosters a dose-dependent regulatory T cell tuned milieu in T1D patients. *Journal of Autoimmunity*, 58, 48–58.

Roweis, S. T. and Saul, L. K., 2000. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290 (5500), 2323–2326.

Roy, S., Lagree, S., Hou, Z., Thomson, J. A., Stewart, R., and Gasch, A. P., 2013. Integrated Module and Gene-Specific Regulatory Inference Implicates Upstream Signaling Networks. *PLoS Computational Biology* [online], 9 (10). Available from: https://www-ncbi-nlm-nih-gov.gate2.inist.fr/pmc/articles/PMC3798279/ [Accessed 26 Jan 2017].

Russell, S., Norvig, P., and Intelligence, A., 1995. A modern approach. *Artificial Intelligence. Prentice-Hall, Egnlewood Cliffs* [online], 25. Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.259.8854&rep=rep1&type=pdf [Accessed 9 Jan 2017].

Ruths, T., Ruths, D., and Nakhleh, L., 2009. GS2: an efficiently computable measure of GO-based similarity of gene sets. *Bioinformatics*, 25 (9), 1178–1184.

Saeed, A. I., Sharov, V., White, J., Li, J., Liang, W., Bhagabati, N., Braisted, J., Klapa, M., Currier, T., Thiagarajan, M., Sturn, A., Snuffin, M., Rezantsev, A., Popov, D., Ryltsov, A., Kostukovich, E., Borisovsky, I., Liu, Z., Vinsavich, A., Trush, V., and Quackenbush, J., 2003. TM4: a free, open-source system for microarray data management and analysis. *BioTechniques*, 34 (2), 374–378.

Sammut, C. and Webb, G. I., 2011. *Encyclopedia of Machine Learning*. Springer Science & Business Media.

Sandifer, C. E., 2007. *The Early Mathematics of Leonhard Euler*. MAA.

Sanger, F., Nicklen, S., and Coulson, A. R., 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74 (12), 5463–5467.

Sarode, S. C., Anand, R., Sarode, G. S., and Patil, S., 2016. Somatic Mutation Theory/Tissue Organization Field Theory: Has the Premise been Wrong All along? *World Journal of Dentistry*, 7 (4), 167–168.

Sarver, A. L., 2010. Toward Understanding the Informatics and Statistical Aspects of Micro-RNA Profiling. *Journal of Cardiovascular Translational Research*, 3 (3), 204–211.

Schek, H.-J. and Pistor, P., 1982. Data Structures for an Integrated Data Base Management and Information Retrieval System. *In*: *VLDB* [online]. Citeseer, 197–207. Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.102.4812&rep=rep1&type=pdf [Accessed 30 Dec 2016].

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O., 1995. Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. *Science*, 270 (5235), 467–470.

Scientific American, 1890. THE NEW CENSUS OF THE UNITED STATES-THE ELECTRICAL ENUDRATING MECHANISM. *Scientific American*, 30 August 1890, p. 132.

Segal, E., H. Wang, et D. Koller. 2003. « Discovering molecular pathways from protein interaction and gene expression data ». *Bioinformatics* 19 (suppl_1): i264-72. doi:10.1093/bioinformatics/btg1037.

Serbu, R., Marza, B., and Borza, S., 2016. A Spatial Analytic Hierarchy Process for Identification of Water Pollution with GIS Software in an Eco-Economy Environment. *Sustainability*, 8 (11), 1208.

Sewell, W., 1964. MEDICAL SUBJECT HEADINGS IN MEDLARS. *Bulletin of the Medical Library Association*, 52, 164–170.

Shadbolt, N., Berners-Lee, T., and Hall, W., 2006. The semantic web revisited. *IEEE intelligent systems*, 21 (3), 96–101.

Shannon, C. E., 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27, 379–423.

Shendure, J. and Ji, H., 2008. Next-generation DNA sequencing. *Nature Biotechnology*, 26 (10), 1135–1145.

Shows, T. B., McAlpine, P. J., Boucheix, C., Collins, F. S., Conneally, P. M., Frézal, J., Gershowitz, H., Goodfellow, P. N., Hall, J. G., Issitt, P., Jones, C. A., Knowles, B. B., Lewis, M., McKusick, V. A., Meisler, M., Morton, N. E., Rubenstein, P., Schanfield, M. S., Schmickel, R. D., Skolnick, M. H., Spence, M. A., Sutherland, G. R., Traver, M., Van Cong, N., and Willard, H. F., 1987. Guidelines for human gene nomenclature. An international system for human gene nomenclature (ISGN, 1987). *Cytogenetics and Cell Genetics*, 46 (1–4), 11–28.

Sirin, E., Parsia, B., Grau, B. C., Kalyanpur, A., and Katz, Y., 2007. Pellet: A practical OWL-DL reasoner. *Web Semantics: Science, Services and Agents on the World Wide Web*, 5 (2), 51–53.

Skarstad, K., Steen, H. B., and Boye, E., 1983. Cell cycle parameters of slowly growing Escherichia coli B/r studied by flow cytometry. *Journal of Bacteriology*, 154 (2), 656.

Smith, B., 2003. Ontology. *In*: *Blackwell Guide to the Philosophy of Computing and Information* [online]. Wiley-Blackwell, 155–166. Available from: http://philpapers.org/rec/SMIO-2 [Accessed 3 Jan 2017].

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J., Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.-A., Scheuermann, R. H., Shah, N., Whetzel, P. L., and Lewis, S., 2007. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, 25 (11), 1251–1255.

Smith, B. J., Darzins, P. J., Quinn, M., and Heller, R. F., 1992. Modern methods of searching the medical literature. *The Medical Journal of Australia*, 157 (9), 603–611.

Smith, E. S., 1993. On the Shoulders of Giants: From Boole to Shannon to Taube: The Origins of Computerized Information from the Mid-19th Century to the Present. *Information Technology and Libraries*, 12 (2), 217.

Smyrnaios, Nikos, et Smyrnaios Franck. 2009. « L'actualité selon Google L'emprise du principal moteur de recherche sur l'information en ligne ». *Communication & langages* 2009 (160): 95-109. doi:10.4074/S0336150009002087.

Sokolova, M., Japkowicz, N., and Szpakowicz, S., 2006. Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. *In*: *Australasian Joint Conference on Artificial Intelligence* [online]. Springer, 1015–1021. Available from: http://link.springer.com.gate2.inist.fr/chapter/10.1007/11941439_114 [Accessed 25 Jan 2017].

Soper, H. E., Young, A. W., Cave, B. M., Lee, A., and Pearson, K., 1917. On the Distribution of the Correlation Coefficient in Small Samples. Appendix II to the Papers of 'Student' and R. A. Fisher. *Biometrika*, 11 (4), 328–413.

Soto, A. and Sonnenschein, C., 2008. Pathologie: l'exemple du cancer. *In: Biologie du XXIe siècle : évolution des concepts fondateurs* [online]. De Boeck Supérieur, 299–326. Available from: https://books.google.tn/books?id=lbdZ-Hz7zGAC&pg=PA311&lpg=PA311&dq=la+th%C3%A9orie+de+l%27organisation+des+tissus+(TOFT)&source=bl&ots=Qak1ULPgte&sig=gPak1-kz8oOt0Q92CYf7PKbWgUk&hl=fr&sa=X&redir_esc=y#v=onepage&q=TOFT&f=false.

Souza-e-Silva, H., Savino, W., Feijóo, R. A., and Vasconcelos, A. T. R., 2009. A Cellular Automata-Based Mathematical Model for Thymocyte Development. *PLOS ONE*, 4 (12), e8233.

Spengler, S. J., 2000. Bioinformatics in the Information Age. *Science*, 287 (5456), 1221–1223.

Springer, T. A., 1990. Adhesion receptors of the immune system. *Nature*, 346 (6283), 425–434.

Stambouli, N., Dridi, M., Wei, N.-N., Jlizi, A., Bouraoui, A., and Elgaaied, A. B. A., 2014. Structural insight into the binding complex: β-arrestin/CCR5 complex. *Journal of Biomolecular Structure & Dynamics*, 32 (6), 866–875.

Stein, L., 2001. Genome annotation: from sequence to biology. *Nature Reviews Genetics*, 2 (7), 493–503.

Stelzer, Gil, Naomi Rosen, Inbar Plaschkes, Shahar Zimmerman, Michal Twik, Simon Fishilevich, Tsippi Iny Stein, et al. 2016. « The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses ». Current Protocols in Bioinformatics 54 (juin): 1.30.1-1.30.33. doi:10.1002/cpbi.5.

Stroeymeyt, N., Jordan, C., Mayer, G., Hovsepian, S., Giurfa, M., and Franks, N. R., 2014. Seasonality in communication and collective decision-making in ants. *Proceedings of the Royal Society B: Biological Sciences* [online], 281 (1780). Available from: https://www-ncbi-nlm-nih-gov.gate2.inist.fr/pmc/articles/PMC4027394/ [Accessed 27 Dec 2016].

Sturtevant, A. H., 1913. The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association. *Journal of Experimental Zoology*, 14 (1), 43–59.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P., 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102 (43), 15545–15550.

Susskind, A. M., Schwartz, D. F., Richards, W. D., and Johnson, J. D., 2005. Evolution and Diffusion of the Michigan State University Tradition of Organizational Communication Network Research. *Communication Studies*, 56 (4), 397–418.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K. P., Kuhn, M., Bork, P., Jensen, L. J., and von Mering, C., 2015. STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, 43 (Database issue), D447–D452.

Talbi, B. and Bessière, P., 1991. A parallel genetic algorithm for the graph partitioning problem. *In*: [online]. hal. Available from: https://hal.inria.fr/file/index/docid/89203/filename/Talbi91b.pdf [Accessed 17 Dec 2016].

Tan, A.-H., 1999. fouille des textes: The state of the art and the challenges. *In*: *Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases* [online]. 65–70. Available from: http://www.ntu.edu.sg/home/asahtan/papers/tm_pakdd99.pdf [Accessed 13 Jan 2017].

Tasgin, M., Herdagdelen, A., and Bingol, H., 2007. Community Detection in Complex Networks Using Genetic Algorithms. *arXiv:0711.0491 [physics]* [online]. Available from: http://arxiv.org/abs/0711.0491 [Accessed 15 Dec 2016].

Teacher, A. G. F., Griffiths, D. J., Hodgson, D. J., and Inger, R., 2013. Smartphones in ecology and evolution: a guide for the app-rehensive. *Ecology and Evolution*, 3 (16), 5268–5278.

Thapa, D. R., Tonikian, R., Sun, C., Liu, M., Dearth, A., Petri, M., Pepin, F., Emerson, R. O., and Ranger, A., 2015. Longitudinal analysis of peripheral blood T cell receptor diversity in patients with systemic lupus erythematosus by next-generation sequencing. *Arthritis Research & Therapy*, 17, 132.

Thomas-Vaslin, V., 2015. Complexité multi-échelle du système immunitaire: Evolution, du chaos aux fractales. *In*: *Le vivant critique et chaotique* [online]. Edition Matériologiques, 333-. Available from: https://www.researchgate.net/publication/281558174_Complexite_multi-echelle_du_systeme_immunitaire_Evolution_du_chaos_aux_fractales [Accessed 14 Feb 2017].

Thomas-Vaslin, V., 2016. Understanding and Modelling the Complexity of the Immune System. *In*: [online]. Presented at the First Complex Systems Digital Campus World E-Conference 2015, Springer, 1–9. Available from: https://www.researchgate.net/publication/299541844_Understanding_and_Modelling_the_Complexity_of_the_Immune_System [Accessed 14 Feb 2017].

Thomas-Vaslin, V., Altes, H. K., Boer, R. J. de, and Klatzmann, D., 2008. Comprehensive Assessment and Mathematical Modeling of T Cell Population Dynamics and Homeostasis. *The Journal of Immunology*, 180 (4), 2240–2250.

Thomas-Vaslin, V., Six, A., Ganascia, J.-G., and Bersini, H., 2013. Dynamical and Mechanistic Reconstructive Approaches of T Lymphocyte Dynamics: Using Visual Modeling Languages to Bridge the Gap between Immunologists, Theoreticians, and Programmers. *Frontiers in Immunology* [online], 4. Available from: http://journal.frontiersin.org/article/10.3389/fimmu.2013.00300/abstract [Accessed 16 Feb 2017].

Turing, A. M., 1936. On computable numbers, with an application to the Entscheidungsproblem. *J. of Math*, 58 (345–363), 5.

Tseng, George C., Min-Kyu Oh, Lars Rohlin, James C. Liao, et Wing Hung Wong. 2001. « Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects ». *Nucleic Acids Research* 29 (12): 2549-57. doi:10.1093/nar/29.12.2549.

Vempati, U. D., Chung, C., Mader, C., Koleti, A., Datar, N., Vidović, D., Wrobel, D., Erickson, S., Muhlich, J. L., Berriz, G., Benes, C. H., Subramanian, A., Pillai, A., Shamu, C. E., and Schürer, S. C., 2014. Metadata Standard and Data Exchange Specifications to Describe, Model, and Integrate Complex and Diverse High-Throughput Screening Data from the Library of Integrated Network-based Cellular Signatures (LINCS). *Journal of Biomolecular Screening*, 19 (5), 803–816.

Venkatesan, A., Kim, J.-H., Talo, F., Ide-Smith, M., Gobeill, J., Carter, J., Batista-Navarro, R., Ananiadou, S., Ruch, P., and McEntyre, J., 2016. SciLite: a platform for displaying text-mined annotations as a means to link research articles with biological data. *Wellcome Open Research*, 1, 25.

Venkatesan, K., Rual, J.-F., Vazquez, A., Stelzl, U., Lemmens, I., Hirozane-Kishikawa, T., Hao, T., Zenkner, M., Xin, X., Goh, K.-I., Yildirim, M. A., Simonis, N., Heinzmann, K., Gebreab, F., Sahalie, J. M., Cevik, S., Simon, C., de Smet, A.-S., Dann, E., Smolyar, A., Vinayagam, A., Yu, H., Szeto, D., Borick, H., Dricot, A., Klitgord, N., Murray, R. R., Lin, C., Lalowski, M., Timm, J., Rau, K., Boone, C., Braun, P., Cusick, M. E., Roth, F. P., Hill, D. E., Tavernier, J., Wanker, E. E., Barabási, A.-L., and Vidal, M., 2009. An empirical framework for binary interactome mapping. *Nature Methods*, 6 (1), 83–90.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Miklos, G. L. G., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Francesco, V. D., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R.-R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z. Y., Wang, A., Wang, X., Wang, J., Wei, M.-H., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S. C., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y.-H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigó, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y.-H., Coyne, M., Dahlke, C., Mays, A. D., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith,

T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A., and Zhu, X., 2001. The Sequence of the Human Genome. *Science*, 291 (5507), 1304–1351.

Cerf, V. C. and Khan, R. E., 1974. A Protocol for Packet Network Intercommunication. *IEEE Transactions on Communications* [online], 22 (5). Available from: http://sites.ijrit.com/papers/sept/V2I938.pdf?attredirects=0 [Accessed 31 Dec 2016].

Von Neumann, J. and Godfrey, M. D., 1945. First Draft of a Report on the EDVAC. *IEEE Annals of the History of Computing*, 15 (4), 27–75.

Wang, C., Yang, S., Sun, G., Tang, X., Lu, S., Neyrolles, O., and Gao, Q., 2011. Comparative miRNA Expression Profiles in Individuals with Latent and Active Tuberculosis. *PLOS ONE*, 6 (10), e25832.

Wang, J. Z., Du, Z., Payattakool, R., Yu, P. S., and Chen, C.-F., 2007. A new method to measure the semantic similarity of GO terms. *Bioinformatics*, 23 (10), 1274–1281.

Watson, J. and Crick, F., 1953. Molecular Structure of Nucleic Acids A Structure for Deoxyribose Nucleic Acid. *Nature*, 171, 737–738.

Weatherall, D. J., 2001. Phenotype—genotype relationships in monogenic disease: lessons from the thalassaemias. *Nature Reviews Genetics*, 2 (4), 245–255.

Weber, M., 2014. Experiment in Biology. *In*: Zalta, E. N., ed. *The Stanford Encyclopedia of Philosophy* [online]. Metaphysics Research Lab, Stanford University. Available from: https://plato.stanford.edu/archives/win2014/entries/biology-experiment/ [Accessed 23 Jan 2017].

Weisberg, S., 2005. *Applied Linear Regression*. John Wiley & Sons.

Williams, M. E., 1988. Defining Information Science and the Role of ASIS. *Bulletin of the American Society for Information Science*, April 1988, pp. 17–18.

Wilson, Stephen j., Angela d. Wilkins, Chih-hsu Lin, Rhonald c. Lua, et Olivier Lichtarge. 2016. « Discovery of functional and disease pathways by community detection in protein-protein interaction networks ». Pacific Symposium on Biocomputing. 22: 336-47.

Wong, M. T., Ong, D. E. H., Lim, F. S. H., Teng, K. W. W., McGovern, N., Narayanan, S., Ho, W. Q., Cerny, D., Tan, H. K. K., and Anicete, R., 2016. A high-dimensional atlas of human T cell diversity reveals tissue-specific trafficking and cytokine signatures. *Immunity*, 45 (2), 442–456.

Wu, C. and Zhang, D., 2016. Identification of early-stage lung adenocarcinoma prognostic signatures based on statistical modeling. *Cancer Biomarkers: Section A of Disease Markers*.

Wu, H., Su, Z., Mao, F., Olman, V., and Xu, Y., 2005. Prediction of functional modules based on comparative genome analysis and Gene Ontology application. *Nucleic Acids Research*, 33 (9), 2822–2837.

Wu, Z. and Palmer, M., 1994. Verbs semantics and lexical selection. *In*: *Proceedings of the 32nd annual meeting on Association for Computational Linguistics* [online]. Association for Computational Linguistics, 133–138. Available from: http://dl.acm.org/citation.cfm?id=981751 [Accessed 19 Dec 2016].

Xuning, L., Genshan, Z., Weihua, C., Hong, T., and Liying, D., 2012. Reaserch on agricultural information retrieve based on ontology. *In*: *Advances in Future Computer and Control Systems*. Berlin: Springer Science & Business Media, 153–157.

Yang, Y., Han, L., Yuan, Y., Li, J., Hei, N., and Liang, H., 2014. Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. *Nature Communications*, 5, 3231.

Yu, H., Gao, L., Tu, K., and Guo, Z., 2005. Broadly predicting specific gene functions with expression similarity and taxonomy similarity. *Gene*, 352, 75–81.

Zeigler, J. F. (zig), 1997. Gutenberg, the Scriptoria, and Websites. *Journal of Scholarly Publishing*, 29 (1), 36–43.

Zhang, L., Zhou, W., Velculescu, V. E., Kern, S. E., Hruban, R. H., Hamilton, S. R., Vogelstein, B., and Kinzler, K. W., 1997. Gene Expression Profiles in Normal and Cancer Cells. *Science*, 276 (5316), 1268–1272.