Université de Montréal

Un dictionnaire pour faciliter la recherche des gènes dans la littérature et sur Internet

par
Amine Halawani

Département d'informatique et de recherche opérationnelle
Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de Maître ès sciences (M.Sc.)
en informatique

Mai, 2007

Université de Montréal
Faculté des études supérieures

Ce mémoire intitulé:

Un dictionnaire pour faciliter la recherche des gènes dans la littérature et sur Internet

présenté par:

Amine Halawani

a été évalué par un jury composé des personnes suivantes:

Guy Lapalme
président-rapporteur

François Major
directeur de recherche

Jian-Yun Nie
membre du jury

Mémoire accepté le 27 juin 2007

RÉSUMÉ

Les gènes sont beaucoup manipulés en bioinformatique. Cette popularité s'explique par la nature des gènes. Chaque gène représente une unité de l'information génétique qui est transmise d'un individu à sa descendance. L'information génétique regroupe tous les caractères d'un individu.

Chaque jour, des centaines d'articles sont publiés avec des références directes ou indirectes sur des gènes. Chacun de ces articles représente un potentiel d'information pour de nombreux chercheurs. Malheureusement, l'état actuel de la formalisation de la biologie fait que cette information n'est pas facilement et directement accessible. Les gènes en particulier n'ont aucune nomenclature officielle et un gène donné peut avoir plusieurs noms ou codes. Par conséquent, il peut être difficile de gérer l'information des gènes par des moyens automatisés.

Ce mémoire présente notre travail dans le domaine des systèmes d'intégration pour la recherche des gènes et de l'information des gènes. Nous proposons une plate-forme appelée Ge-Di (pour *Gene-Dictionary*) pour la recherche de l'information des gènes ainsi que l'extraction d'information des gènes. La plate-forme est divisée en trois modules : le dictionnaire, le moteur de recherche basé sur l'algorithme d'*Aho-Corasick* et l'outil *Chromosome Browser* qui utilise l'emplacement chromosomique pour retrouver un gène donné. Le fonctionnement de ces modules est décrit et des exemples des différentes applications de la plate-forme sont donnés.

La plate-forme propose deux nouvelles techniques pour les moteurs de recherche. La première technique introduit une étape intermédiaire dans le scénario typique de recherche : requête-résultats. Sachant qu'un intervalle de temps est perdu en consultant la base de données, nous avons développé un module interactif informant l'utilisateur du nombre de résultats potentiels pour éviter les recherches infructueuses. La deuxième technique a consisté à substituer le type de données généralement utilisé (nom/symbole/identifiant de BD) lors de la recherche des gènes par l'emplacement chromosomique. La plate-forme reprend également l'idée de la suggestion (*did you mean* de Google) lorsque la recherche est infructueuse. Finalement, la plate-forme

peut être étendue à d'autres utilisations. Une interface de programmation (*Application Programming Interface*) a été créée pour que les utilisateurs puissent lancer des requêtes plus complexes ainsi que développer de nouvelles applications.

Mots clés : Informatique, biologie, système d'intègration, dictionnaire, gènes, Aho-Corasick.

ABSTRACT

Genes are widely used in biology. This importance is explained by the definition of a gene. Each gene represents a unit of the genetic information transmitted to offspring. The genetic information determines the appearance and behavior of a person. Hundreds of articles are published daily with direct or indirect references to genes. Each article is a potential piece of information for many researchers. However, the current state of formalism in biology makes this information difficult to rearch. Genes in particular do not have an official nomenclature and a given gene can appear under different names in articles. It is therefore difficult to handle gene information using automated means.

This thesis presents our work in the field of system integration for gene search and gene information extraction. We propose a platform called Ge-Di (for *Gene-Dictionary*) as well as new tools for text mining. The platform is divided into three modules : the search engine based on the algorithm of *Aho-Corasick*, the dictionary and *Chromosome Browser* that uses the chromosomal location to find a given gene. The modules and application samples of the platform are introduced. The platform proposes two new techniques for search engines. The first technique introduces an intermediate step in the typical search scenario: query-results. Since a time period is lost while querying the database, an estimation has been developed to inform the user about the potential number of results to avoid unsuccessful searches. The second technique substitutes the usual input (Symbol/Name/ID) by the chromosomal location. The platform also uses the suggestion idea (*Did you mean ?*) taken from the *Google* search engine, when the query is not successful. Finally, the platform can be extended. An Application Programming Interface has been implemented to allow the user to perform complex queries as well as to develop new applications.

Keywords: integration systems, computer science, biology, dictionary, genes, Aho-Corasick.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES APPENDICES

(dédicace) à mes grands-parents....

REMERCIEMENTS

L'auteur souhaite remercier son directeur de recherche François Major pour sa direction dans le domaine difficile de la bioinformatique, Marc Parisien et Martin Larose pour leur expertise technique, Emmanuelle Permal, Nicolas Saint-Onge, Sébastien Christin et Mohamed Tikah Marrakchi pour leurs nombreux conseils.

CHAPITRE 1

INTRODUCTION

Le sujet de recherche traité dans ce mémoire est bidisciplinaire. Il s'agit de résoudre un problème en biologie par des moyens informatiques. L'introduction du mémoire est divisée en trois parties. Dans la première partie, nous introduisons certaines notions biologiques telles que la transcription et la traduction. Ces deux notions sont essentielles à la compréhension du mémoire et des *micro acides ribonucléiques* (*microARN*[1]). Dans la deuxième partie, nous présentons une recherche de relations d'expression entre gènes dans la littérature ainsi que les difficultés rencontrées lors de cette approche. Cette recherche permettrait de découvrir de nouveaux microARN ou d'expliquer certaines relations d'expression avec des microARN connus. Finalement, dans la troisième partie, nous proposons une approche pour résoudre ces difficultés.

1.1 Contexte Biologique

Les mots soulignés sont définis dans le lexique disponible à l'annexe I.

1.1.1 La transcription et la traduction

Les gènes sont des éléments centraux de la cellule. Chaque gène représente une unité de l'information génétique qui est transmise d'une cellule mère à une cellule fille. Au niveau de l'organisme, l'information génétique regroupe les caractéristiques de chaque individu comme la couleur des yeux, la forme du visage, etc. chez l'humain.

Un gène est précisément défini comme étant une séquence d'acide désoxyribonucléique (ADN) codant un seul ARN messager (ARNm), qui code une protéine (Figure

[1]Petites molécules importantes au développement d'une cellule.

1.1) [Wik07a]. Les protéines sont essentielles à une cellule car elles lui permettent d'assurer ses fonctions essentielles comme la respiration et l'assimilation des nutriments. Le passage de l'ADN à la protéine est un processus complexe composé de deux grandes étapes : la transcription et la traduction.

La transcription est un copiage de certaines parties d'un des deux brins de l'ADN. Ce copiage est catalysé par une enzyme appelée ARN polymérase. Cette dernière détermine quelles sont les régions de l'ADN à copier grâce à des signaux parsemés le long de l'ADN. Le produit de la transcription est un ARN (transcrit primaire) qui est une copie presque identique de l'un des deux brins de l'ADN (la thymine de l'ADN est remplacée par l'uracile dans l'ARN). Il existe une étape intermédiaire entre la transcription et la traduction qui est l'épissage. Les ARN qui sont issus de la transcription, ne sont pas forcement traduits, dans ce cas, on parle d'ARN non codant.

La traduction est le phénomène qui suit la transcription pour les ARN codants (Figure 1.1). Elle fait appel au ribosome qui va permettre de lire l'information de l'ARNm (transcrit primaire devenu ARNm dans le cytoplasme). Ainsi chaque groupe de 3 nucléotides (dit codon) de l'ARNm va correspondre à un acide aminé dans la protéine naissante. La formation de la protéine naissante requiert une molécule d'ARN de transfert (ARNt). Cette molécule permet le transfert de l'acide aminé à la protéine naissante. Une fois cette lecture terminée, la protéine adopte une structure tridimensionnelle et accomplit sa tâche spécifique.

En général, on dit qu'un gène s'exprime lorsque la protéine correspondante est produite. Les biologistes peuvent également mesurer le taux d'expression des gènes et définir leur niveau d'expression normal. Une telle mesure permet de déceler quels gènes sont impliqués dans une maladie donnée. Par exemple, Le gène *CCND1* est surexprimé dans le cancer du sein [HA98]. Cette surexpression confère à *CCND1* un rôle important dans ce type de cancer. Les biologistes s'intéressent également aux relations qui peuvent se tisser entre les expressions de différents gènes. Ainsi, la sur-expression d'un gène A peut être accompagnée d'une sous-expression du gène B. Cette *relation d'expression* est tout particulièrement importante si le gène B

est responsable d'une fonction essentielle de la cellule. Dans le cancer des ovaires par exemple, le gène *STAT1* est sur-exprimé tandis que le gène *THBS1* est sous-exprimé [LMBK+07]. Étudier cette relation est donc nécessaire pour comprendre les mécanismes du cancer des ovaires. Ces relations restent cependant très hypothétiques. La différence dans le taux d'expression des deux gènes peut être une coïncidence ou un processus complexe composé de plusieurs étapes intermédiaires. C'est pour cette raison qu'une grande partie de ces relations demeurent inexpliquées. Les miARN que nous présentons dans la section suivante peuvent contribuer à expliquer certaines de ces relations.

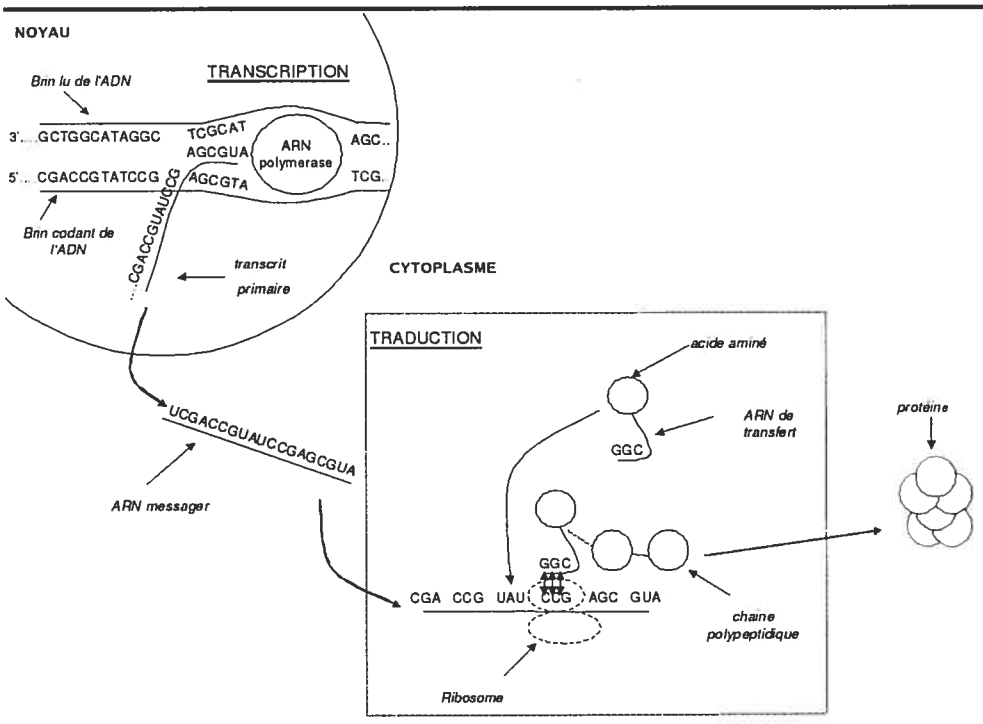

FIG. 1.1 – **Transcription de l'ADN et traduction de l'ARNm :** L'ADN est retranscrit en transcrit primaire dans le noyau à l'aide de l'ARN polymérase puis l'ARN messager (transcrit primaire mature) est traduit en une chaîne d'acides aminés dans le cytoplasme à l'aide du ribosome et des ARN de transfert.

1.1.2 Les microARN

Dans le paragraphe précédent, nous avons discuté des ARN non-codants qui sont connus depuis longtemps. Par exemple, le ribosome (voir le bas de la figure 1.1) est un complexe ribonucléoprotéique, composé de molécules d'ARN et de protéines. Récemment, un rôle actif dans la régulation de la traduction a été découvert pour les ARN non-codants [LFA93] : les *miARN* (ou microARN) [ABB+03].

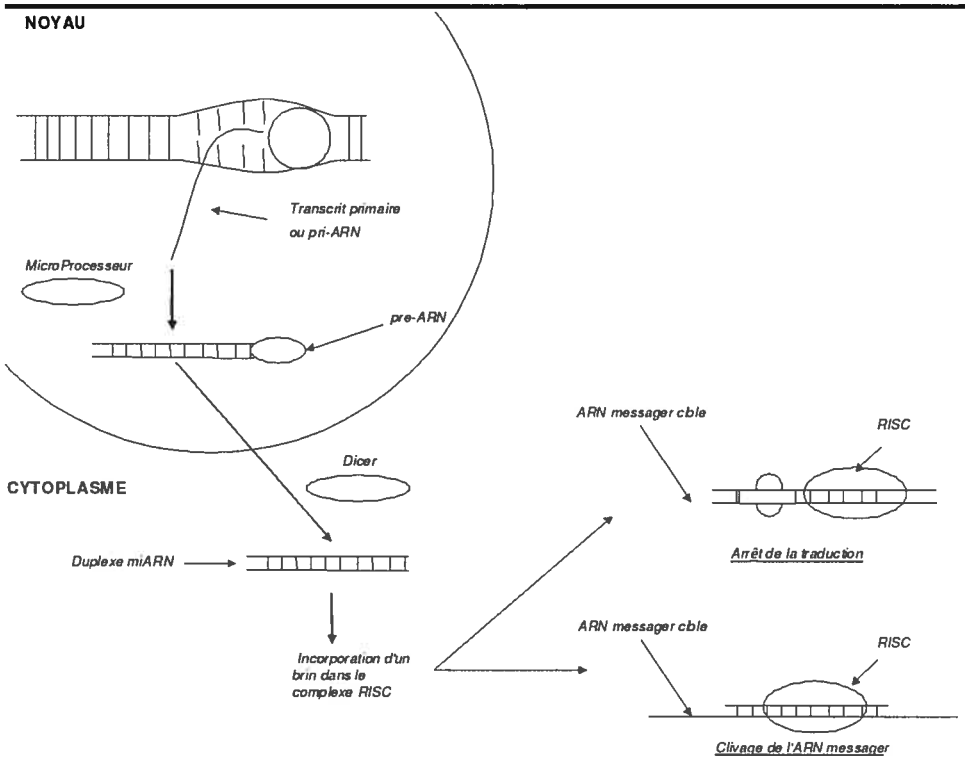

FIG. 1.2 – **La génération des miARN :** La production des miARN est différente de celle des protéines. En premier lieu, il y a transcription de l'ADN en transcrit primaire (comme pour la production des protéines). Le transcrit primaire passe ensuite par plusieurs étapes où il est modifié avant de se combiner avec le complexe RISC dans les eucaryotes.

La production des miARN partage certaines similarités avec la production des protéines (Figure 1.2). Comme pour la production des protéines, l'ADN est re-

transcrit en transcrit primaire. Ce dernier est transformé en une tige boucle par un complexe enzymatique (microprocessor[2]) dans le noyau puis envoyé dans le cytoplasme. Dans le cytoplasme, la tige boucle est modifiée par la molécule *Dicer* pour devenir un ARN double brin appelé *le duplexe miARN*. Finalement, un des brins est incorporé à un ensemble de protéines : le complexe RISC. Le complexe RISC confère à l'ARN non-codant des propriétés fonctionnelles. Le complexe RISC peut venir se coller à des ARN messagers. Cependant, ce phénomène dépend de la complémentarité des bases azotées (Tableau 1.1).

Séquences complémentaires :
5'— AGCUGA — 3' (*mARN*)
3'— UCGACU — 5' (*microARN*)
Séquences non complémentaires :
5'— AG**C**C**G**A — 3' (*mARN*)
3'— U**A**G**A**CU — 5' (*microARN*)

TAB. 1.1 – **Complémentarité des bases azotées dans l'ARN :** Pour que la séquence d'ARNm et la séquence de microARN soient complémentaires, il faut que chaque *adénine* (A) corresponde à un *uracile* (U) et que chaque *guanine* (G) corresponde à une *cytosine* (C). Ainsi les deux premières séquences sont complémentaires mais ce n'est pas le cas pour les deux autres.

Les biologistes ont remarqué néanmoins deux cas spécifiques. Lorsque la complémentarité est parfaite entre les deux séquences, il y a clivage de l'ARNm, c'est-à-dire que l'ARNm va être automatiquement détruit par la cellule. Lorsque la complémentarité n'est pas parfaite (c'est-à-dire certains couples de nucléotides ne respectent pas les appariements possibles), l'ARNm n'est pas détruit, mais il ne sera pas traduit, par conséquent la protéine correspondante ne sera pas produite non plus. Ainsi, les miARN sont impliqués dans la régulation de l'expression génétique par leur présence (interruption de la production de la protéine) ou leur absence

[2]Le complexe enzymatique est formé de l'enzyme *Drosha* et de la protéine *DGCR8* [KK07].

(production de la protéine). Les cibles des miARN sont spécifiques à leur séquences. Ces deux cas ont valu un grand intérêt des biologistes pour les miARN à cause des nombreuses retombées y compris dans le domaine des relations d'expression.

1.2 La recherche de relations d'expression dans la littérature

Dans la section précédente, nous avons fait une distinction entre la production d'une protéine et la production d'un miARN. Récemment, certains biologistes ont découvert que pour certains gènes, les deux processus pouvaient être confondus [RGJAB04]. Cependant, ce phénomène s'applique uniquement aux organismes eucaryotes. Chez les organismes eucaryotes, les gènes codants sont constitués d'une suite d'*exons* et d'*introns*. Les exons sont les parties du transcrit primaire qui vont former l'ARNm et les introns sont les parties du transcrit primaire qui vont être enlevées lors de l'épissage. Les biologistes ont découvert que certains miARN pouvait exister dans les introns et par conséquent, la production d'une protéine pouvait être accompagnée de celle d'un miARN.

Cette découverte nous a tout particulièrement intéressé car elle introduisait une nouvelle perspective pour la recherche de miARN. Cette perspective implique les relations d'expression des gènes. Nous avons vu dans la section 1.1.1, que certaines relations d'expression pouvaient se tisser entre les gènes. Dans notre cas, nous nous sommes intéressés aux relations où une expression normale (ou sur-expression) d'un gène A est accompagnée d'une sous-expression d'un gène B. Nous avons déjà mentionné que ces relations sont pour la plupart hypothétiques. Dans notre hypothèse, le gène A code une protéine et un miARN. Ce dernier empêche la traduction de l'ARNm de B d'où la sous-expression de B (Figure 1.3).

Cette perspective permettrait soit de trouver de nouveaux miARN, soit d'expliquer des relations d'expression par des miARN connus. Avant de pouvoir appliquer cette approche, nous devons trouver les relations d'expression. Deux démarches s'offrent à nous dans ce cas. La première démarche est d'aller rechercher ces relations dans

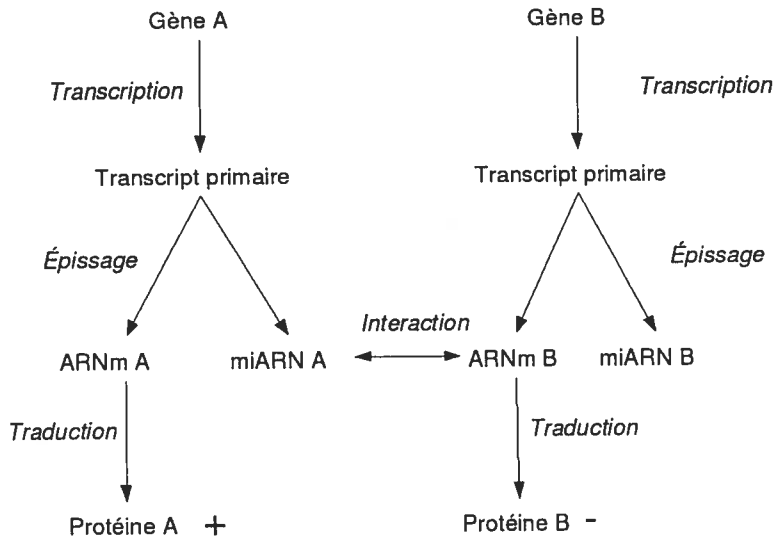

FIG. 1.3 – **Nouvelle Perspective pour la recherche des miARN :** Étant donné que les miARN peuvent se retrouver dans les introns, un gène peut coder en même temps une protéine et un miARN. Par conséquent, une sur-expression ou expression normale d'un gène A peut induire une sous-expression du gène B. Cette régulation peut être causée par un miARN contenu dans les introns de A. A noter que les étapes de la production des miARN ont été omises dans cette figure par besoin de clarté.

les bases de données. Certains bases de données telles que *Stanford Microarray*[3] *Database* [DBG+07] permettent d'extraire des relations d'expression. La plupart des informaticiens opteraient pour cette démarche. La deuxième démarche est de rechercher les relations dans la littérature (articles et publications surtout). Cette démarche est typique du biologiste.

Nous avons opté pour la démarche d'un biologiste car nous pensons trouver dans la littérature une quantité brute d'informations. Dans le domaine biomédical, la littérature scientifique est considérée comme la source principale de connaissances [Wre04] [Koe05]. Les bases de données utiles existent mais elles ne couvrent pas toujours l'information des publications scientifiques [TPM+99] [Koe05].

[3] *Microarray ou puce à ADN est une expérience qui permet de mesurer le niveau d'expression de milliers de gènes à la fois.*

Notre processus de recherche s'est donc résumé à quatre étapes :

1. Trouver des relations d'expression de souris dans la littérature
2. Prétraiter ces relations. Ne conserver que celles où l'expression normale (ou la surexpression) d'un gène A a été accompagnée d'une sous-expression d'un gène B.
3. Extraire les introns de la séquence A
   - si un ou plusieurs miARN sont connus dans les introns de A
     (a) Tester si une complémentarité est possible entre un des miARN de A et l'ARNm de B à l'aide du programme *miranda*[4] [EJG+03]
   - si aucun miARN n'est connu dans les introns de A
     (a) Trouver si des miARN peuvent exister dans les introns de A à l'aide d'un outil de détection de miARN tel que *MiRAlign*[5] [WZL+05]
     (b) Tester si une complémentarité est possible entre un des miARN potentiels de A et l'ARNm de B à l'aide de *miranda*

Malheureusement, ce processus de recherche n'a pas pu aller à son terme. A chaque étape du processus, des difficultés sont survenues. Dans la première étape, beaucoup de relations sont rejetées car les symboles utilisés par l'auteur ne correspondent pas à des gènes. Dans les autres étapes, plusieurs recherches dans des bases de données et dans la littérature ont été effectuées pour retrouver la séquence génomique ou la séquence de l'ARN messager de certains gènes.
Ces différentes difficultés nous ont poussé à voir si elles ne pouvaient pas être évitées. Nous avons directement songé à une automatisation du processus. Cependant, avant de pouvoir appliquer une automatisation, il faut comprendre les raisons des difficultés. Nous nous sommes alors intéressés à la recherche d'information dans le domaine biomédical.

[4]http ://www.microrna.org/
[5]http ://bioinfo.au.tsinghua.edu.cn/miralign/

1.3 Recherche d'information dans le domaine biomédical

La recherche d'information consiste à selectionner des documents pouvant contenir des informations pertinentes à une requête. Elle consiste également à rechercher l'information dans des bases de données [Wik07b]. Le moteur de recherche *Google* est un exemple d'outil de recherche d'information. Les hommes d'affaires portent particulièrement une grande attention aux outils de recherche d'information dans la prise de décision. Ces outils sont utilisés pour faire des analyses de marchés en vue de lancer un nouveau produit, analyser des valeurs boursières en vue d'investir dans une autre compagnie ou encore pour sélectionner des candidats pour les offres d'emploi. Les deux outils les plus connus dans le monde des affaires sont *Text Miner*[6] de *SAS* et *Clementine*[7] de *SPSS Inc.*

Les biologistes et les bioinformaticiens se sont également intéressés aux outils de recherche d'information pour permettre de traiter plus d'articles scientifiques en moins de temps. La plupart des articles publiés dans le domaine biomédical sont référencés dans *MEDLINE* [Wik07c][8]. À la rédaction du mémoire, approximativement 15 millions d'articles sont contenus dans *MEDLINE* avec une croissance annuelle de 400,000 articles [LH05]. Il devient évident qu'un chercheur ne peut pas lire un millier d'articles par jour et que des moyens informatiques sont nécessaires pour que celui-ci soit au courant des découvertes scientifiques dans son domaine de recherche [XFH+07] [Wre04] [JSB06] [ESB06] [Koe05].

Malheureusement, l'application des méthodes informatiques n'a pas eu le succès escompté. En général, en faisant un prétraitement sur un ensemble de documents, on peut construire un ensemble avec des documents homogènes (par exemple sélectionner que des textes en français). Cet ensemble devient plus facile à traiter qu'un ensemble qui contient à la fois des documents en français et des documents en anglais [WSVM+03] car le tout ne contient qu'un seul langage. Cette règle ne s'applique pas aux articles biomédicaux où un tel prétraitement ne permet pas de

[6] *http ://www.sas.com/technologies/analytics/datamining/textminer/index.html*
[7] *http ://www.spss.com/clementine/*
[8] http ://www.ncbi.nlm.nih.gov/entrez/query.fcgi ?DB=pubmed

faciliter le traitement subséquent. Le langage biomédical est à l'origine très ambigu. Cette ambiguïté empêche les algorithmes de recherche d'information d'être directement appliqués. Avant de pouvoir extraire une information pertinente sur un gène, il faut pouvoir identifier le gène. C'est une tâche qui paraît simple mais qui ne l'est pas [LH05]. L'identification d'une entité biomédicale (particulièrement les gènes) dans un texte est maintenant considérée comme le goulot d'étranglement pour la recherche d'information dans le domaine biomédical [ESB06] [NBD+06].

1.4 L'identification d'une entité biomédicale dans un texte

Le problème de la reconnaissance d'une entité biomédicale dans la littérature est assez connu et un certain nombre d'articles le soulignent [CLF05] [LH05] [ESB06] [XFH+07] [WSVM+03].
Pour expliquer la difficulté à identifier un nom de gène dans un texte, il faut comprendre qu'une entité biologique dans un texte peut correspondre à :

1. un ou plusieurs gènes pouvant provenir d'une (*HOX4*) ou plusieurs espèces (*CDC21*, *p53*) [FZ06] [CLF05]
2. une protéine (*tumor protein p53*[9]), une cellule ou un virus [ESB06].
3. une maladie (*cancer*, *diabetes*)
4. un terme de la langue anglaise (*was*, a) [FZ06] [CLF05] ou à une date (*DEC1* [ZRK+04])

De plus, les appellations de gènes ne sont pas soumises à une stricte vérification par une autorité[10] quelconque. Par conséquent :

- Les auteurs sont libres de nommer un gène comme ils le souhaitent (*Pokemon*[11])
- Un gène peut changer plusieurs fois de noms (*BCRA2* s'appelait *FAS1*)

[9] Le gène et la protéine correspondante ont le même nom.
[10] La section 1.5 présente néanmoins deux comités responsables d'assigner à chaque gène un symbole et un nom unique.
[11] ©*Nintendo* a finalement forcé les auteurs à changer le nom du gène [Sim05].

Ces changements rendent difficiles l'identification des gènes dans un texte donné mais également la recherche d'articles concernant un gène donné. Dans le premier cas, l'utilisateur doit connaître le gène en question sinon il y a de grandes chances qu'il pense qu'il s'agit du dessin animé. Dans le deuxième cas, toute recherche pour de l'information sur *BCRA2* doit être accompagnée d'une recherche simultanée sur *FAS1* mais encore faut-il que l'utilisateur sache que *BCRA2* s'appelait *FAS1*. Plusieurs études ont été effectuées pour essayer de déterminer l'ampleur du problème. Nous nous sommes intéressés à deux de ces études en particulier.

La première étude [CLF05] a examiné les noms de gènes des organismes eucaryotes dans les publications. Les auteurs de l'étude ont remarqué que 74.7 % des appellations de gènes mentionnées dans les articles étaient des synonymes. Or la plupart des bases de données tel que *Entrez Gene* ou *MGI*[12] utilisent le symbole officiel ou le nom officiel[13] pour identifier et classer les gènes et ne contiennent pas toujours toutes les appellations possibles pour un gène. Les symboles et noms officiels représentent respectivement 17.7% et 7.6% des gènes mentionnés dans les articles.

La deuxième étude est plus complexe. Elle a été effectuée par Katrin Fundel et Ralf Zimmer [FZ06]. Ces derniers ont compilé cinq dictionnaires provenant de la levure, la drosophile, la souris, le rat et l'humain. Chaque dictionnaire contient les noms des gènes et de protéines pour chacun des organismes. Dans cette étude, comme la plupart des études (ou outils) faites sur le sujet, aucune différence n'est faite entre les noms des protéines et ceux des gènes. Fundel et Zimmer ont analysé l'ambiguïté entre :

- Les noms de gènes et protéines au sein d'une même espèce
- Les noms de gènes et protéines de deux espèces
- Les noms de gènes et protéines d'une espèce et les mots de la langue anglaise
- Les noms de gènes et protéines d'une espèce et les mots du vocabulaire biomédical

[12]Ces deux bases de données contenant l'information des gènes seront présentées ultérieurement dans le mémoire.

[13]un symbole ou un nom de gène est officiel lorsqu'il a été approuvé par les comités présentés dans la section 1.5

Pour l'ambiguïté dans la même espèce, la première place revient à l'humain (3.24%) donc le génome humain contient beaucoup de noms de gènes qui correspondent à plus d'un seul gène (*homonyme*[14]). La souris fait partie des organismes ayant le moins d'homonymes (2.00%).

Pour l'ambiguïté entre espèces, le couple humain-souris possède le degré d'ambiguïté le plus élevé (20.00%) c'est-à-dire que le génome humain et celui de la souris possèdent un grand nombre de gènes qui partagent le même nom.

Pour l'ambiguïté avec les termes de la langue anglaise, la drosophile dépasse la plupart des autres organismes (2.78%) y compris la souris (0.75%) et l'humain (1.00%).

Pour l'ambiguïté avec les termes du vocabulaire biomédical, la drosophile est encore une fois en tête de liste (0.95%) mais la souris (0.45%) et l'humain (0.90%) suivent de très près.

Cette étude est probablement la plus complète sur le sujet. Les auteurs malheureusement n'analysent pas vraiment leurs résultats et ne font que constater le problème. Néanmoins, il faut souligner qu'il y a eu restriction de l'environnement dans chacun des 4 cas (chaque expérience d'ambiguïté était une comparaison entre deux dictionnaires). Cette restriction explique des pourcentages peu élevés mais lors de la lecture d'un article en langue anglaise, cette restriction n'existe pas et le lecteur fait face à au moins deux sources d'ambiguïté : l'ambiguïté avec la langue anglaise et l'ambiguïté avec le vocabulaire biomédical.

La première étude souligne que tout outil de détection de gènes (dans un texte donné) doit être accompagné ou relié à une base de données contenant l'information des gènes. La raison en est simple. Éviter d'avoir un nom de gène identifié qui ne peut pas être exploité directement (c'est-à-dire que son information doit être recherchée dans d'autre bases de données ou dans la littérature). Dans ce cas, on peut essayer de rechercher le nom officiel du gène mais cette tâche manuelle est coûteuse et difficile y compris pour des biologistes [LH05].

[14]Un homonyme est un mot qui a plusieurs sens. Pour les gènes, un homonyme est un nom qui correspond à plusieurs gènes.

La deuxième étude quant à elle souligne le besoin d'un dictionnaire des noms et symboles de gènes connus pour la détection des gènes dans la littérature [FZ06] [SMWK06]. Il y a actuellement trop d'ambiguïtés avec la langue anglaise ou le vocabulaire biomédical, pour que la détection du gène se fasse sur la base du contexte ou de la syntaxe des phrases seulement [ESB06]. Et puisque la première étude nécessite un lien direct avec une base de données, le dictionnaire pourrait être tiré de la même base de données.

1.5 Les démarches de formalisation

Dans la section précédente, nous avons mentionné des noms et des symboles officiels de gènes. Dans cette section, nous expliquons la définition des symboles et noms officiels et les comités responsables de leur attribution.

Les démarches pour *mettre de l'ordre* dans les noms de gènes existent depuis longtemps. Pour l'humain, la démarche de formalisation la plus connue est celle du *Human Genome Nomenclature Comittee* (HGNC ou HUGO) qui a commencé au début des années 80 [SAB+79]. Le but de ce comité est d'assigner un nom et un symbole unique à chaque gène en suivant certaines directives [WLDP02] [WBL+02]. Les directives stipulent que le symbole d'un gène humain doit être strictement composé de lettres majuscules et de chiffres arabes. Le symbole ainsi formé et son nom correspondant sont dits *officiels* et les autres appellations sont considérées comme des *synonymes* ou *alias*.

Chez la souris, une démarche similaire existe. Il s'agit du MGI (*Mouse Genome Informatics*). Le but est similaire à HGNC mais le symbole d'un gène de la souris doit avoir la première lettre majuscule et les suivantes en minuscules. Depuis quelques années, MGI et HGNC tentent d'assigner le même symbole et nom aux gènes qui sont considérés similaires dans les deux espèces[15]. Cette mesure peut expliquer le degré élevé d'ambiguïté entre la souris et l'humain dans l'étude de Fundel

[15]Le symbole doit respecter cependant les directives concernant les lettres majuscules dans chacune des deux espèces.

et Zimmer (voir la section 1.4).
À la composition de ce mémoire, plus de 24,000 gènes humains sont passés par ce processus. Malheureusement, l'utilisation de cette nomenclature n'est pas assez stricte et n'est pas renforcée. Les directives des comités sont remises en cause par certains auteurs [Lud06]. Par exemple, Le gène *LFNG* (*lunatic fringe homolog (Drosophila)*) n'a aucun rapport avec la folie (*lunacy*). Les auteurs se retrouvent donc libres d'utiliser les noms comme bon leur semble et tout particulièrement dans leurs articles scientifiques. Des analyses vont dans ce sens. Les auteurs continuent à utiliser les anciennes ou leur propre nomenclature [CLF05] [WSVM+03].

1.6 Notre approche

Suite à notre recherche des causes des difficultés, nous avons décidé d'explorer le domaine de la linguistique dans l'espoir de trouver des outils existants ou des méthodes utiles. Dans ce domaine, il existe une branche (la *morphologie*) qui étudie la structure interne des mots dans une langue donnée. Nous nous sommes alors demandés si les appellations et en particulier les symboles pouvaient avoir une structure particulière. Nous avons par conséquent analysé les symboles des gènes humains pour extraire des caractéristiques propres[16]. Malheureusement, cette caractérisation n'a fait que confirmer la grande diversité des symboles de gènes et l'absence de caractéristiques propres.
Suite à cette déconvenue, nous avons décidé d'exploiter les connaissances acquises lors de la recherche des causes des difficultés. La plupart des articles mentionnent des outils pour la recherche de l'information des gènes. Ainsi, nous avons mesuré la performance et les avantages des outils tel que *GAPSCORE* [CSA04] et *Google* [PBMW98]. Malheureusement, ces deux outils ne permettent pas un lien direct vers les bases de données. Nous avons alors examiné les bases de données suivante : *MGI* [EBK+05], *HGNC* [WLDP02] et *Entrez Gene* [MOPT05]. Il se peut très bien que ces bases de données auraient pu proposer des outils pouvant nous

[16]Cette caractérisation sera présentée dans le chapitre 1.

aider. Malheureusement, nous n'avons pas trouvé d'outils adoptés à notre besoin. Cependant, Entrez Gene s'est trouvé être une base de données assez complète et avec des options utiles permettant le développement d'outils.
De cette évaluation des différents outils, nous avons réalisé que l'information des gènes existe (Entrez Gene), mais que son accès nécessite un certain travail ce qui peut ralentir les requêtes qui comprennent plus qu'un gène. Il fallait développer un outil qui se basait sur Entrez Gene mais en même temps plus précis et mieux structuré que ce dernier. C'est ainsi qu'est venue l'idée de la plate-forme *Ge-Di*. Nous avons appelé la plate-forme Ge-Di pour *Gene Dictionary*. Elle permet de définir un gène (son symbole, son nom, son emplacement chromosomique), et de présenter ses alias (ses différentes appellations) et finalement prouver son appartenance à un organisme tout comme un dictionnaire peut définir un mot dans un langage donné et prouver son appartenance à ce dernier.

1.7 Présentation de la plate-forme

Ge-Di[17] est une plate-forme qui est composée de deux moteurs de recherche. Le premier moteur de recherche permet de retrouver un gène à partir de :

- son symbole officiel
- son nom officiel
- un de ses alias

Le premier moteur de recherche est basé sur l'algorithme d'*Aho-Corasick* [Aho75]. Il est composé de deux modes de recherche : exact et approximatif. Le mode exact peut être assimilé à une recherche exacte de l'entrée dans une base de données. Le mode approximatif quant à lui tente de pousser la similarité de caractères au maximum entre l'entrée et certains termes de la base de données. Il retourne les noms et symboles qui partagent un degré élevé de similarité. Ce moteur de recherche est également accompagné d'outils pour faciliter l'expérience de la recherche et parfois accélérer l'accès à l'information. Le premier outil est l'estimation en temps réel

[17]http ://www-lbit.iro.umontreal.ca/GD/

du nombre exact de résultats. L'estimation s'effectue en même temps que l'utilisateur tape sa requête et lui permet de contrôler le volume généré de résultats. Le deuxième outil est déclenché lorsque la requête de l'utilisateur a échoué et suggère des noms et symboles de gène qui partagent un grand degré de similarité avec l'entrée.

Le deuxième moteur de recherche permet de retrouver un gène à partir de son emplacement chromosomique. C'est une alternative au premier moteur de recherche mais également un moyen d'extraire la liste des gènes présents dans un emplacement chromosomique donné.

Une fois le gène trouvé, la plate-forme définit le gène et permet d'accéder à l'information du gène sur des bases de données externes (Entrez Gene, MGI ou HGNC[18]) ainsi que les séquences de l'ARNm sur la base de données *Entrez Nucleotide*[19] [WBB+07].

Finalement, la plate-forme contient également une interface de programmation qui facilite l'extraction de l'information de plusieurs gènes en même temps ainsi que le développement d'outils à base de la plate-forme.

1.8 Survol du mémoire

Ce mémoire est divisé en six parties. Dans la première partie, nous évaluons les méthodes et les systèmes utiles pour la recherche de l'information dans le domaine biomédical. Dans la deuxième partie, nous discutons du moteur de recherche principal basé sur l'algorithme d'Aho-Corasick. Nous poursuivons sur l'outil *Chromosome Browser* qui peut être considéré comme une alternative au moteur de recherche principal. Dans la troisième partie, nous parlons de la base de données et nous nous attardons sur les moyens mis en place pour la consistance de l'information. Dans la quatrième partie, nous présenterons quelques applications pour l'interface graphique et l'interface de programmation. Cette présentation est ac-

[18]Les bases de données HGNC et MGI sont spécifiques à un organisme tandis que Entrez Gene englobe plusieurs organismes.

[19]http ://www.ncbi.nlm.nih.gov/entrez/query.fcgi ?db=Nucleotide

compagnée de tests de comparaison avec des systèmes existants. Finalement dans la cinquième partie, nous abordons les différentes améliorations et modifications qui pourraient être effectuées sur la plate-forme.

Ce mémoire est accompagné d'un lexique dans l'annexe I et de l'article sur la plate-forme dans l'annexe II. L'article sera prochainement soumis au journal *bio-informatics* (*Oxford Journals*).

CHAPITRE 2

MÉTHODES ET SYSTÈMES UTILES

2.1 Caractérisation des symboles humains

Après avoir analysé nos difficultés, nous avons décidé d'explorer une approche inspirée d'une branche de la linguistique. Certains linguistes étudient la structure interne des mots dans une langue donnée. Il se peut très bien que les symboles de gènes aient une structure interne caractéristique. Nous avons donc analysé les symboles des gènes humains pour extraire des caractéristiques propres (tableau 2.1). Nous avons mesuré la longueur minimale, longueur maximale, nombre de chiffres, nombre de lettres pour chacun des 32,816 symboles du génome humain. Nous avons également mesuré la distance de *Levenshtein* [Lev65]. Cette dernière mesure le nombre minimal d'opérations (insertion, substitution et suppression) pour passer d'un mot à un autre. C'est également un indicateur du degré de similarité entre deux mots donnés. Les 32,816 gènes connus de l'humain ont été téléchargés du site ftp de la base de données Entrez Gene. Pour la distance de Levenshtein, nous avons choisi de ne pas mesurer la distance entre chacun des gènes (32816 !) mais plutôt de calculer la distance pour les 1000 premiers symboles avec le reste de la liste (32,816,000 distances). Malheureusement, les résultats de cette expérience n'ont fait que confirmer la grande diversité des symboles de gènes et l'absence de caractéristiques propres. Nous avons quand même tenté d'analyser chacune des variables mesurées à commencer par la longueur des symboles. Cette dernière ne permet pas de définir un intervalle de longueur, elle souligne plutôt deux catégories de symboles : courts et longs. Pour le contenu, il n'y a pas proprement dit un consensus, le nombre de lettres et chiffres y est très variable. Finalement, la distance de Levenshtein prouve qu'il n'existe non plus un nombre d'opérations moyen pour passer d'un symbole à un autre. Le nombre moyen d'opérations nous semble trop élevé pour être considéré comme une caractéristique utile au développement

Longueur	min max moy.	3 19 8.55
Nombre de chiffres	min max moy.	0 14 2.12
Nombre de Lettres	min max moy.	0 8 2.18
Distance Levenshtein n=32816000	min max moy.	1 17 6.0

TAB. 2.1 – **Résultats de l'analyse syntaxique des symboles des gènes humains :** Les 32816 symboles du génome humain ont été analysés. Pour la distance de Levenshtein (qui mesure le degré de similarité entre deux chaînes de caractères), la distance a été calculée pour les 1000 premiers gènes de la liste d'où n= 1000 × 32816.

d'un outil de détection.

Suite à cette déconvenue, nous avons décidé d'opter pour une nouvelle approche. Cette dernière utilise les connaissances acquises lors de la recherche des causes de nos difficultés (sections 1.3 et 1.4). La plupart des articles mentionnent des outils pour la recherche de l'information des gènes. Nous avons décidé d'évaluer ces outils en espérant trouver celui qui nous permettrait de faire avancer notre recherche.

2.2 Outils de détection de gènes

Naturellement, notre attention s'est portée sur les outils de détection de gènes dans un texte donné. Ces outils permettent une lecture assez rapide[1] d'un grand nombre d'articles et l'extraction des gènes mentionnés (étape 1 de la recherche de relations d'expression - section 1.2). L'outil le plus intéressant que nous avons trouvé est GAPSCORE[2] [CSA04]. Il est basé sur un algorithme d'apprentissage machine. Malheureusement, comme tout outil de recherche d'information, GAPS-

[1]La comparaison se fait avec un traitement manuel

[2]http ://bionlp.stanford.edu/gapscore/

CORE génère des faux positifs. Il est vrai que les faux positifs ne peuvent pas être réduits totalement. Certains projets peuvent tolérer un grand nombre de faux positifs tandis que d'autres pas [LH05]. Pour les besoins de notre recherche, nous avons besoin d'un outil qui réduise les faux positifs pour éviter de rechercher un gène qui n'existe pas (GAPSCORE ne fait pas de lien direct avec une base de données). Si un gène n'existe pas dans une base de données, il se peut très bien que GAPSCORE se soit trompé ou que la base de données ne contienne pas l'information sur ce gène. Dans ce cas, un choix doit être fait entre rejeter l'entité biologique et aller tenter de la rechercher dans d'autres bases de données. Ce choix est facile si le nombre d'entités à vérifier est peu nombreux mais dès lors qu'il devient élevé, l'avantage d'automatiser cette étape est perdu. De plus, GAPSCORE ne fait pas la différence entre les noms de gènes et les noms de protéines. Dans la plupart des bases de données l'information du gène permet d'accéder à l'information de la protéine mais le contraire n'est pas toujours vrai. Il ne faut pas oublier qu'il existe des interactions protéines-protéines. Ces dernières sont très étudiées dans le domaine pharmaceutique. Par conséquent, il est préférable d'exclure les protéines des relations d'expression afin d'éviter d'expliquer un phénomène déjà connu.
Nous avons essayé de trouver des outils similaires mais à notre grande surprise, GAPSCORE est le seul programme disponible publiquement [MR04]. Nous avons alors songé à certains moteurs de recherche. Ils ne permettent pas de retrouver un gène dans un texte donné mais ils peuvent potentiellement identifier un terme biomédical donné (c'est-à-dire déterminer si XYZ est un gène). Notre choix s'est tout naturellement porté vers le plus connu d'entre eux : *Google*.

2.3 Google

Le moteur de recherche le plus connu actuellement sur Internet est probablement Google[3]. Le mot google a même fait son apparition dans le dictionnaire *Merriam-*

[3]http ://www.google.com

Webster[4] pour signifier une recherche sur Internet.

Google repose sur un système de classement [PBMW98]. Plus une page est référencée par d'autres pages plus sa position est élevée dans le classement. Cette promotion est d'autant plus forte si les pages de référence ont des positions élevées dans le classement. Pour la recherche des gènes, Google ne s'avère pas être aussi efficace. Si *CDC21*, *Cdc21* et *cdc21* représentent respectivement un gène humain, un gène de la souris et un gène pouvant appartenir soit à *Pyrococcus abyssi GE5*[5] soit à *Archaeoglobus fulgidus DSM 4304*[6]. Pour Google, ces trois gènes sont équivalents et le moteur de recherche va ignorer les lettres majuscules et les convertir en lettres minuscules. Si GAPSCORE permet une réponse à la question *"est ce que CDC21 appartient à la souris ?"*[7], le moteur de recherche Google ne permet pas une réponse directe. Pour la requête *Cdc21*, le moteur de recherche Google ne permet pas de déterminer l'appartenance de *Cdc21* au génome de la souris. Cette insensibilité à la casse s'explique par la volonté de Google de générer un grande nombre de résultats quelque soit la position du mot recherché dans la phrase ou de son orthographe. Ce choix donne à Google son efficacité lors de la recherche en général sur Internet mais le rend malheureusement peu adapté à la recherche des gènes.

De plus, dans leur article [PBMW98], les concepteurs de Google précisent que les pages dont les adresses mentionnent *cgi-bin* ne sont pas référencées par Google or la plupart des bases de données pour les gènes (*Entrez Gene*, *MGI* et *HGNC*) utilisent des pages dynamiques. Néanmoins en pratique, cette mesure ne semble plus être appliquée à la rédaction du mémoire. Le dynamisme du contenu des pages sur Internet est sans doute pris en charge par le nombre important de serveurs : 450,000 [Car06].

---

[4]http ://www.m-w.com/

[5]un microbe vivant dans les profondeurs de la mer

[6]un microbe vivant dans les gisements d'huile ou les sources d'eau chaude

[7]La réponse n'est pas nécessairement véridique.

2.4 Les bases de données

Après avoir essayé respectivement un outil de détection de gènes dans un texte et un moteur de recherche, nous nous sommes tournés vers les bases de données de gènes. Ces dernières sont très importantes pour l'étape 3 (section 1.2). Il se peut très bien que ces bases de données proposent des outils pouvant nous aider. Nous avons commencé notre évaluation par des bases de données spécialisées avant de nous tourner vers une base de données générale. Un certain nombre de bases de données spécialisées dans les gènes existent sur Internet. Nous avons choisi celles reliées aux démarches de formalisation. Nous pensons y trouver des outils plus adaptés à l'état actuel des gènes.

HGNC[8] et **MGI**[9] : HGNC et MGI ne sont pas seulement des comités chargés de gérer la nomenclature des gènes (section 1.5). Ce sont également deux bases de données publiques contenant l'information des gènes (Nom et symbole officiel, alias, emplacement chromosomique, liens vers la séquence génomique et ARNm) spécialisées respectivement dans l'humain et la souris. Nous avons pensé que MGI et HGNC étaient les deux outils qui nous permettraient de continuer la recherche et de l'accélérer. Malheureusement notre espoir était vain. Il n'y pas de lien entre ces deux bases de données. Une appellation donnée dans la littérature peut bien représenter un gène chez l'humain et un gène chez la souris. Pour confirmer ou infirmer cette possibilité, l'utilisateur doit simultanément consulter les deux bases de données. Ce cheminement est très coûteux dès que le nombre de gènes à vérifier est grand. Ces deux bases de données ont cependant des avantages importants :

- accès direct aux séquences génomiques et ARNm si elles sont connues
- possibilité de téléchargement du contenu de la base de données[10]

Toutefois, ces deux outils possèdent un moteur de recherche très flexible. La flexibilité n'est pas un problème en soi mais lors d'une recherche précise, un grand nombre de résultats est généré et l'utilisateur doit en faire le tri. À titre d'exemple,

[8] http ://www.gene.ucl.ac.uk/nomenclature/
[9] http ://www.informatics.jax.org/
[10] Cette option s'applique uniquement à HGNC.

la recherche de *p53* dans HGNC et MGI donne respectivement 92 et 72 gènes. Actuellement, il n'y aucun moyen de régler la précision du moteur de recherche (c'est-à-dire éviter que la requête soit seulement "**p53**" où '*' représente 0 ou plus de caractères). Il y a également un autre aspect qui nous a déplu. HGNC ne contient que l'information des gènes approuvés par son comité, c'est-à-dire approximativement 24,000 gènes. Or ce nombre est nettement inférieur au nombre de gènes connus (environ 40,000). Les bases de données spécialisées ne nous ont donc pas entièrement satisfaits. Nous nous sommes alors tournés vers les bases de données générales. Comme pour les moteurs de recherche sur Internet, un outil se démarque des autres. Il s'agit de Entrez Gene.

Entrez Gene[11] : Pour éviter d'avoir à faire des recherches simultanées sur deux bases de données, nous nous sommes tournés vers les bases de données plus générales englobant plus qu'un seul organisme. Notre attention s'est portée surtout sur les bases de données du *National Center for Biotechnology Information* (NCBI) de l'Institut National de Santé des États-Unis (http ://www.ncbi.nlm.nih.gov/). Le site est considéré par beaucoup de biologistes comme la référence pour l'information des gènes. De plus, *Entrez Gene* qui est une des bases de données du NCBI, importe de manière régulière le contenu des bases de données de HGNC et MGI. Entrez Gene ne contient pas seulement l'information basique des gènes (nom et symbole officiel, emplacement chromosomique, alias). Elle contient également les séquences génomiques et d'ARNm dans sa base de données connexe Entrez Nucleotide[12].
Malheureusement, la flexibilité est également présente. S'il peut sembler logique que les moteurs de recherche de HGNC et MGI ne soient pas sensibles à la casse, le moteur de recherche Entrez Gene ne peut pas vraiment se permettre de ne pas l'être. Comme nous l'avons déjà mentionné pour Google, *CDC21*, *Cdc21* et *cdc21* sont des gènes différents provenant de 3 organismes distincts. Entrez Gene ne fait pas également la distinction entre ces trois gènes. Cependant, l'interface permet de restreindre les espèces pour la recherche mais la précision ne peut pas être réglée.

[11]http ://www.ncbi.nlm.nih.gov/entrez/query.fcgi ?db=gene
[12]http ://www.ncbi.nlm.nih.gov/entrez/query.fcgi ?db=nucleotide

Du point de vue contenu, Entrez Gene contient tous les gènes connus de l'humain et la souris. Son contenu peut être téléchargé.

Lors de cette évaluation des différents outils, nous avons réalisé que l'information des gènes existe (Entrez Gene) mais qu'un certain travail doit être effectué pour y accéder. C'est ainsi que nous avons eu l'idée de développer notre propre outil pour permettre d'identifier un gène et d'accéder rapidement à son information. Cet outil est basé sur Entrez Gene mais son moteur de recherche est plus précis que celui de Entrez Gene.

CHAPITRE 3

LES OUTILS DE RECHERCHE

Dans les deux premiers chapitres, nous avons souligné nos difficultés lors de la recherche d'information des gènes dans les bases de données à cause du grand nombre de résultats générés et de l'absence de moyens pour contrôler ce nombre. Dans ce chapitre, nous présentons nos démarches pour développer un moteur de recherche plus adapté à nos besoins. Avant de présenter nos démarches, nous allons introduire une autre difficulté que nous avons découverte lors de l'élaboration du moteur de recherche.

3.1 Mémoriser un nom de gène

Nous savons que dans le domaine de la recherche, le besoin d'information est constant pour valider les travaux en cours ou même une future expérience à effectuer. Le type d'information peut être très varié mais si nous considérons que toute recherche peut être assimilée à une question à laquelle nous voulons répondre : Il existe alors deux types de réponses : courtes (*Est ce que p53 appartient au génome de la souris ?*) et longues (*comment p53 est-il impliqué dans le cancer ?*). À la rédaction du mémoire, la plupart des moteurs de recherche tentent de répondre aux questions du deuxième type. C'est pour cela que nous avons souligné que la flexibilité de Entrez Gene n'est pas un problème en soi, et que cela dépend de la question à laquelle on tente de répondre. Rechercher un gène similaire à *p53* ou partageant la même fonction cellulaire peut être facilement accompli à l'aide de Entrez Gene. Mais si nous tentons de confirmer l'appartenance d'un gène à un organisme quelconque, un certain travail doit être effectué.

Il faut également souligner qu'il existe un nombre impressionnant de gènes : À la rédaction du mémoire, le monde scientifique connaît approximativement 40,000 gènes chez l'humain et 60,000 gènes chez la souris. Un biologiste ne peut pas

connaître par coeur tous les gènes d'un organisme donné et il sera forcement difficile de les mémoriser. Cette remarque est valable également dans les domaines spécialisés tel que le cancer (Tableau 3.1). Dans ce domaine, on a dénombré une centaine de gènes pouvant être impliqués directement ou indirectement [FCM+04]. Encore une fois, nous avons remarqué que les biologistes et les bioinformaticiens ne connaissaient pas précisément tous les noms de gènes.

ABI1, ABL2, ACSL6, AF1Q, AF5Q31, AKT1, ARNT, ASPSCR1, ATF1, ATIC, BCL10, BFHD, BIRC3, BMPR1A, BTG1, CBFA2T1, CBFA2T3, CBFB, CCND1, CDC2, CDK4, CHIC2, CHN1, COPEB, COX6C, CTNNB1, CYLD, DDB2, DDIT3, DEK, EIF4A2, EPS15, ERCC2, ERCC3, ERCC5, ERG, ETV4, ETV6, EWSR1, EXT1, EXT2, FANCC, FANCG, FGFR1OP, FGFR3, FH, FIP1L1, FUS, GAS7, GATA1, GMPS, GOLGA5, GPC, GPHN, HIST1H4I, HRAS, HSPCA, IL21R, IRF4, KRAS2, LASP1, LCP1, LHFP, LMO2, LYL1, MADH4, MLF1, MLH1, MLLT3, MLLT6, MNAT1, MSF, MSH2, MSN, MUTYH, MYC, NCOA4, NF2, NPM1, NRAS, PAX8, PCBD, PDGFB, PIM1, PLK2, PNUTL1, POU2F1, PPARG, PRCC, PRKACB, PRKAR1A, PTEN, PTPN11, RABEP1, RAD51L1, RAP1GDS1, RARA, RB1, RET, RHOH, RPL22, SBDS, SDHB, SEPTIN6, SET, SH3GL1, SS18L1, SSX1, SSX2, SSX4, STAT3, TAF15, TCF12, TCL1A, TFE3, TFEB, TFG, TFPT, TFRC, TNFRSF6, TP53, TPM3, TPM4, TRIP11, VHL, WAS, WT1, ZNF198, ZNF278, ZNF384, ZNFN1A1

TAB. 3.1 – **Liste de symboles de gènes pouvant être impliqués dans le cancer :** Il existe une centaine de gènes humains qui sont connus pour avoir un rôle potentiel dans le cancer. Cette liste n'est probablement pas complète mais elle souligne la difficulté à mémoriser tous les noms de gènes pour un biologiste y compris dans un domaine spécialisé tel que le cancer [FCM+04].

Cependant notre expérience a démontré que les biologistes ou les bioinformaticiens arrivaient quand même à se rappeler des noms de gènes en utilisant une méthode assez simple. Cette méthode consiste à trouver dans la chaîne de caractères, une sous-chaîne facilement identifiable (Tableau 3.2). La sous-chaîne peut être une abréviation fortement utilisée par le biologiste, un caractère qui se répète ou encore un chiffre. Cependant, il est rare que deux utilisateurs trouvent la même sous-chaîne pour une chaîne donnée de caractères. Nous avons également remarqué

que la sous-chaîne se trouvait presque toujours soit au début soit à la fin de la chaîne de caractères. Nous avons tenté de trouver une base scientifique pour cette dernière remarque en orthophonie mais malheureusement nous n'avons trouvé aucun ouvrage à ce propos. Nous pensons toutefois que le lecteur cherche automatiquement à lire les symboles de gènes or ces derniers ne peuvent pas être toujours lus (par exemple *FANCG*, doit-on prononcer fan, fank, fansje ?). Le lecteur tente alors de lire ce qui peut être lu et énumérer ce qui ne le peut pas (FAN-C-G). Pour l'occurrence des sous-chaînes au début ou à la fin de la chaîne, nous pensons que la longueur peut être une explication plausible. Les symboles de gènes sont assez courts, il est par conséquent, rare de pouvoir diviser le mot en plus que deux parties (FAN / C-G ou TRIP / 11).

Symbole de gène	**Partie caractéristique**
ABI1	AB (deux premières lettres de l'alphabet)
ARNT	ARN (abréviation de acide ribonucléique)
CCND1	CC (abréviation de Centimètre Cube)
DDB2	DD
FANCG	FAN (un fan est un admirateur d'un artiste)
HIST1H4L	HIST (diminutif de histoire)
PDGFB	PDG (Abréviation de président directeur de général)
TAF15	TAF(travail en argot)

TAB. 3.2 – **Méthode de mémorisation approximative des symboles de gène :** Ce tableau montre que certains symboles de gènes peuvent être mémorisés en recherchant une partie caractéristique facilement identifiable. Les parties caractéristiques sont très subjectives et sont presque toujours différentes d'un individu à un autre.

En règle générale, les moteurs de recherche permettent de retrouver des gènes en utilisant cette méthode d'approximation. Cependant toute mémorisation ou approximation peut contenir des erreurs c'est à dire que le mot mémorisé peut être incorrectement orthographié. Dans ce cas, les moteurs de recherche ne sont plus aussi efficaces. La raison de cet échec est simple, la plupart des moteurs tentent

de retrouver toutes les appellations de gènes qui contiennent *précisément* l'entrée. La nouvelle génération de moteurs de recherche commence à développer des outils qui supposent une possibilité d'erreurs dans la requête ("do you mean" de *Google*). Malheureusement, aucune des bases de données que nous avons mentionnées dans l'introduction ne propose un outil aussi efficace que celui de Google.
Connaissant les spécifications de la plate-forme, nous nous sommes dirigés vers la construction de cette dernière. Avant de débuter la préparation d'un plat cuisiné, il faut trouver les ingrédients les plus appropriés. Dans notre cas, il faut déterminer quel algorithme et quelle structure de données seraient les plus appropriés pour notre projet. En général, lorsqu'il s'agit de rechercher un mot donné dans un ensemble de mots ou dictionnaire, les informaticiens pensent directement à un type particulier de structures de données : les arbres. En bioinformatique, deux arbres sont particulièrement populaires : l'arbres des suffixes et l'arbre d'Aho-Corasick.

3.2 Arbre des suffixes

L'arbre des suffixes est une structure de données qui a été inventée par Weiner en 1973 [Wei73]. L'algorithme initial a été simplifié successivement par Mccreight en 1976 [McC76] et Ukkonen en 1995 [Ukk95]. L'arbre des suffixes permet une représentation d'un ensemble de mots et de leurs suffixes respectifs (Figure 3.1). Pour chaque mot *abcd* du dictionnaire, les suffixes de *abcd* (c'est-à-dire *abcd*, *bcd*, *cd*, *d*) sont présents dans l'arbre et représentent un chemin de la racine[1] à une feuille[2] donnée.
L'arbre des suffixes est assez populaire en bioinformatique pour ses applications telles que la recherche de palindrome ou de répétitions dans les séquences génomiques. L'arbre des suffixes est également connu pour la rapidité de son algorithme de recherche. La recherche d'un mot donné de longueur n prend un temps $O(n)$. Cependant, l'insertion est assez coûteuse et la raison en est simple : Pour chaque mot à

[1]Une racine dans un arbre est un noeud n'ayant aucun arc entrant.
[2]Une feuille dans un arbre est un noeud qui ne possède aucun arc sortant

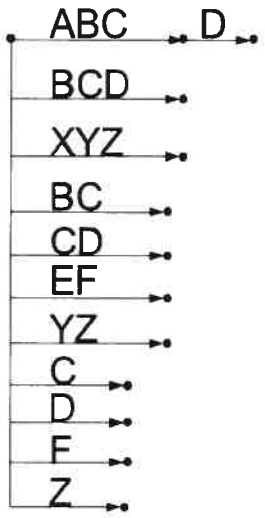

FIG. 3.1 – **Arbre des suffixes :** Le dictionnaire utilisé par l'algorithme contient les mots suivants : ABC, ABCD, EF et XYZ. Chaque suffixe des mots précédents est représenté par un chemin de la racine à une feuille donnée. Un point représente un sommet et une flèche représente un arc. Une racine est un sommet n'ayant aucun arc rentrant et une feuille est un sommet n'ayant aucun arc sortant.

insérer, il faut insérer chacun de ses suffixes. Si le mot à insérer est de longueur n alors la complexité de l'insertion de ce mot va être $O(n+n-1+n-2+n-3+....+1) = O(\frac{n(n+1)}{2}) = O(n^2)$

Cette complexité combinée à l'espace requis nous a fait réfléchir. Nous avons décidé de mesurer les avantages et inconvénients de l'utilisation des arbres des suffixes dans notre projet.

Avantages :

- Rapidité de la recherche
- Grande flexibilité de la recherche (permet de retrouver un mot par un de ses suffixes)

Inconvénients :

- Lenteur de la construction de l'algorithme original [McC76, Ukk95]
- Espace requis [DNM00]

- Aucune étude sur les conséquences de l'optimisation de l'espace sur l'algorithme de recherche
- Un outil additionnel est requis pour retrouver à quel mot appartient un suffixe donné dans l'arbre

A partir de cette liste, il devient évident que les arbres de suffixes ne sont peut-être pas le meilleur outil. De plus, avoir à insérer tous les suffixes de chaque mot du dictionnaire nous semble excessif. Avons-nous besoin de tous les suffixes pour la recherche ? La réponse à la question est non et la méthode que nous avons présentée dans la section 3.1 en est une preuve. De plus, il n'y a pas de méthodes directes pour faire une estimation du nombre potentiel de résultats à part faire plusieurs explorations de l'arbre (trop coûteux). Finalement, lorsque la requête échoue, les moyens pour corriger la requête ou suggérer une requête proche (à la manière du "do you mean" de Google) sont également trop coûteux.
A partir de ces réflexions, nous avons décidé d'analyser l'algorithme d'Aho-Corasick. Ce dernier a été utilisé dans la plate-forme. Dans le paragraphe suivant, nous présentons l'algorithme d'Aho-Corasick et les raisons de notre choix.

3.3 Aho-Corasick

L'algorithme d'Aho-Corasick est un algorithme de recherche de chaînes de caractères dans un ensemble de mots, inventé par Alfred V. Aho et Margaret J. Corasick [Aho75]. Comme les arbres des suffixes, l'algorithme d'Aho-Corasick excelle dans la recherche d'un mot donné : pour un mot de taille n , la complexité de la recherche est $O(n)$. Pratiquement, l'algorithme utilise un arbre orienté pour représenter le dictionnaire ou l'ensemble de mots (Figure 3.2). À la différence de l'arbre des suffixes, l'arbre d'Aho-Corasick contient uniquement le mot et non chacun de ses suffixes. Par conséquent, l'insertion est linéaire pour chacun des mots du dictionnaire. Dans le monde de l'informatique, l'algorithme d'Aho-Corasick est connu pour son implémentation dans l'outil de recherche *grep* sur *Unix*. Dans le monde de la bioinformatique, l'algorithme est souvent utilisé pour la détection de

bouts de séquences spécifiques dans le génome (par exemple, rechercher GGGG dans le génome). Comme pour l'arbre des suffixes, nous avons mesuré ses avantages et inconvénients.

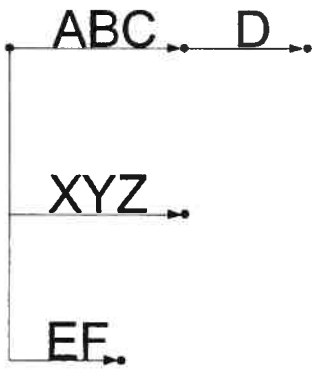

FIG. 3.2 – **Arborescence utilisée par l'algorithme d'Aho-Corasick :** Le dictionnaire utilisé par l'algorithme contient les mots suivants : ABC, ABCD, EF et XYZ. Un point représente un sommet et une flèche représente un arc. Une racine est un sommet n'ayant aucun arc rentrant et une feuille est un sommet n'ayant aucun arc sortant

Avantages :

- Rapidité de la recherche
- Seul une seule copie de chaque mot du dictionnaire existe dans l'arborescence
- Permet facilement l'estimation du nombre de résultats potentiels[3]
- Permet de retrouver un nom de gène en utilisant la méthode illustrée dans la section 3.1[4]

Inconvénients :

- Si la sous-chaîne caractéristique se trouve au milieu du mot, l'algorithme ne permet pas de retrouver le mot

A partir de cette liste, nous avons remarqué un certain potentiel pour l'arbre d'Aho-Corasick et nous avons décidé d'opter pour lui. Dans les paragraphes suivants, nous allons discuter davantage de l'algorithme d'Aho-Corasick.

[3]Cette estimation sera illustrée dans les paragraphes suivants

[4]La recherche en se basant sur cette méthode sera illustrée dans les paragraphes suivants

3.4 Le moteur de recherche basé sur Aho-Corasick

Dans la section précédente, nous avons expliqué notre cheminement avant d'opter pour l'algorithme d'Aho-Corasick. Dans cette section, nous allons expliquer certaines notions nécessaires à la compréhension des outils d'estimation et de suggestion.

Tout dépendant de l'implémentation, les sommets et les arcs dans l'arbre d'Aho-Corasick peuvent avoir différentes fonctions. Dans notre implémentation, nous utilisons, les arcs comme objets de premier ordre. Ils portent des étiquettes qui contiennent l'information. Les sommets, quant à eux, permettent de marquer la fin d'un mot (feuille) ou la fin d'un sous-mot (tout sommet n'étant ni la racine ni une feuille). Chaque mot du dictionnaire est représenté par un chemin de la racine à une feuille donnée[5]. Par conséquent, la recherche d'un mot consiste à explorer l'arborescence. La progression doit être accompagnée de l'identité entre les caractères du mot et ceux des étiquettes. Le mot appartient à l'arbre si et seulement s'il y a eu identité entre tous les caractères des étiquettes et ceux du mot recherché. Néanmoins d'autres conditions peuvent être rajoutées en fonction de l'implémentation et de l'utilisation. L'algorithme de recherche est illustré à la figure 3.3.

La construction de l'arbre d'Aho-Corasick repose sur l'algorithme de recherche. Avant toute insertion, l'algorithme de construction doit déterminer le contenu à insérer ainsi que son emplacement. Le processus de recherche permet de résoudre ces deux inconnues. Deux cas se présentent néanmoins lors de l'insertion. Dans l'algorithme de la figure 3.3, une comparaison est effectuée à la ligne 8 entre chacun des caractères des étiquettes des arcs et ceux du mot recherché. Par conséquent, la recherche d'un mot ne va pas s'arrêter nécessairement sur un sommet. Par exemple, La recherche de *"AC"* dans la figure 3.4 s'arrête sur un arc. L'insertion subséquente diffère selon le lieu d'arrêt de la recherche. Le sommet est sans aucun doute le cas

[5]Similaire aux arbres des suffixes où chaque suffixe représente un chemin de la racine à une feuille donnée

```
Aho Corasick( Entrée : G=(V,A), mot ; sortie : T)
01.début
02.    curseur ← racine
03.    m ← mot
04.    i ← 0
05.   tant que  i = 0
06.       tmp ← m
07.       Pour tout arc a ∈ adjacence { curseur }
08.       si etiquette[a] est un préfixe de m
09.          m ← m - {etiquette[a]}
10.          curseur ← enfant(curseur, etiquette[a])
11.
12.       si tmp = m
13.          T ← 0
14.          i ← 1
15.       sinon
16.          si m = φ et curseur est une feuille
17.             T ← 1
18.          i ← 1
19.fin
```

FIG. 3.3 – **Version simplifiée de l'algorithme d'Aho-Corasick :** Les fonctions **adjacence**, **enfant** et **etiquette** sont des fonctions externes. La fonction **adjacence** permet d'accéder à la liste des arcs d'un sommet. La fonction **etiquette** permet d'accéder à l'étiquette d'un arc. Et finalement, la fonction **enfant** permet d'accéder à un des enfants d'un sommet donné.

le plus simple. Il suffit de vérifier si tous les caractères du mot recherché ont pu être superposés aux caractères des étiquettes. Si c'est n'est pas le cas, alors un arc doit être créé sur ce sommet et son étiquette doit être x (x étant le sous-mot du mot recherché qui n'existe pas dans l'arborescence)(Figure 3.5). Pour le cas d'un arc, une vérification est aussi requise. Si le mot est présent dans l'arborescence alors l'arc doit être brisé en deux arcs pour bien marquer la présence du nouveau mot dans l'arborescence (Figure 3.6). Si le mot est partiellement dans l'arborescence, une cassure est aussi requise mais un arc doit être rajouté au sommet résultant de la cassure (Figure 3.7).

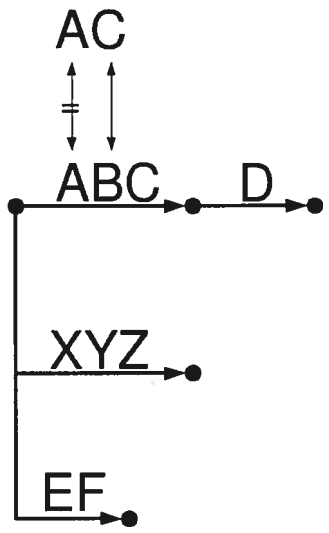

FIG. 3.4 – **Résultat de la recherche de *"AC"* dans l'arborescence :** La première lettre du mot recherché est identique à la première lettre de l'étiquette mais ce n'est pas le cas de la deuxième lettre. Par conséquent, la recherche s'arrête sur un arc. Les flèches à double sens servent à montrer comment se fait la superposition.

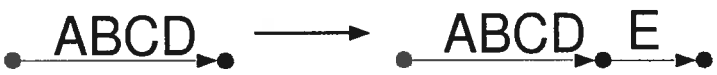

FIG. 3.5 – **Résultat de l'insertion de *"ABCDE"* dans l'arborescence :** Le mot *ABCD* existe auparavant dans l'arbre. En recherchant *ABCDE*, on remarque qu'il existe partiellement dans l'arborescence. Pour marquer sa présence, il suffit d'insérer le caractère de *ABCDE* qui n'est pas contenu dans l'arborescence c'est-à-dire la lettre E.

FIG. 3.6 – **Résultat de l'insertion de *"ABC"* dans l'arborescence :** Le mot *ABCD* existe auparavant dans l'arborescence. En recherchant *ABC*, on remarque que ses trois premiers caractères sont contenus dans l'étiquette de l'arc (*ABCD*). Pour marquer sa présence, il suffit de briser l'arc en deux parties

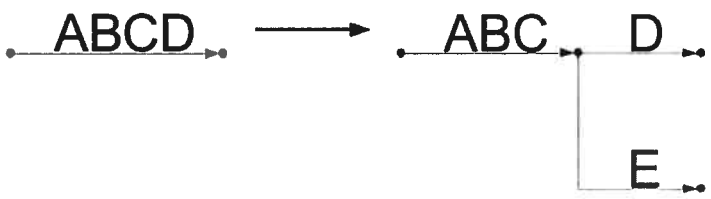

FIG. 3.7 – **Résultat de l'insertion de *"ABCE"* dans l'arborescence :** *ABCD* existe auparavant dans l'arborescence. En effectuant la recherche de *ABCE*, on remarque que la recherche se termine sur un arc et que certains de ses caractères ne sont pas présents dans l'arborescence. Dans ce cas, une cassure et une insertion sont requises.

Maintenant que nous avons défini les notions de bases de l'algorithme d'Aho-Corasick, nous pouvons introduire la suggestion et l'estimation.

3.5 L'estimation et la suggestion du moteur de recherche

Lors de la recherche de l'information des gènes dans la littérature et sur Internet, l'utilisateur doit faire face en général à deux situations avant de pouvoir accéder à l'information souhaitée : trop de résultats à analyser ou pas de résultat. Illustrons ces deux cas par des exemples en allant rechercher l'information des gènes dans la base de données Entrez Gene.

Trop de résultats Supposons que l'utilisateur souhaite obtenir de l'information sur le gène *p53*. Sachant que le gène existe dans le génome humain et le génome de la souris, il limite l'espace de recherche à ces deux espèces puis lance la requête. Il se retrouve avec 787 gènes. S'il a de la patience et du temps, il peut consulter un à un chacun des 787 gènes et retrouver l'information souhaitée. Il peut également essayer de préciser sa requête en rajoutant des mots clefs tel que cancer (p53 est soupçonné d'avoir un rôle important dans le cancer).

Pas de résultats Supposons que l'utilisateur souhaite maintenant retrouver un gène nommé *WAS*. Il tape sa requête et là aucun résultat. À ce stade, il n'a pas

d'autre choix que de changer sa requête. Il peut aller consulter l'article qui mentionnait le gène ou essayer de se rappeler correctement du nom du gène. Dans les deux cas, une nouvelle requête doit être effectuée. En général, l'absence de résultats est un cas très rare.

Cas idéal Supposons qu'il souhaite consulter l'information du gène *CUTA*. Il tape sa requête et il obtient deux résultats : un pour la souris et un pour l'humain. Il consulte l'information recherchée et il a fini.

Les deux premières situations engendrent une perte de temps tout particulièrement si la réponse recherchée est courte (voir section 3.1). Dans le premier cas, un certain nombre de résultats doit être analysé avant de pouvoir retrouver l'information pertinente. Dans le deuxième cas, il doit reformuler plusieurs fois l'information recherchée. Le cas idéal est de trouver directement l'information (c'est-à-dire les premiers résultat sur la page des résultats). Par conséquent, il serait donc intéressant d'être capable de calibrer les résultats c'est à dire de réduire le nombre de résultats à son minimum en utilisant un estimateur exact du nombre de résultats correspondants à la requête. En même temps, il serait souhaitable de pouvoir avoir des suggestions quand la recherche n'est pas fructueuse. Définissons la fonction et l'utilité de ces deux modules.

Estimation exacte du nombre de résultats : Nous avons souligné dans l'introduction ainsi que dans le paragraphe précédent, le grand nombre de résultats qui peut être généré par des outils tel que Google ou Entrez Gene. Une nouvelle génération de moteur de recherche comme *Spotlight*[6] de *MAC OS X* introduit une nouvelle vision du processus de recherche et cette vision est illustrée par la phrase suivante tirée de la page officielle de spotlight[7] :

[6]Spotlight est un outil de recherche de fichiers sur le système d'exploitation *MAC OS X*. Ce programme est équivalent à *Google Desktop Search* sur *Windows*

[7]http ://www.apple.com/macosx/features/spotlight/

"Stop looking. Start finding. With Spotlight, you can find anything on your computer as quickly as you type".

La dernière partie de la phrase est tout particulièrement importante car elle implique que des résultats sont générés et modifiés au fur et à mesure que l'utilisateur tape sa requête. Il suffit de faire une simple comparaison avec le moteur de recherche de *Windows XP* pour réaliser le gain de temps. Avec un tel outil, l'attente des résultats est négligeable par rapport à l'attente des résultats avec les moteurs de recherche traditionnels. Cette estimation exacte des résultats permet d'arrêter de taper la requête dès que l'information recherchée a été trouvée mais également de modifier la requête lorsque les résultats ne sont pas adéquats. Ce cheminement permet à l'utilisateur de maîtriser les résultats d'une manière explicite.

Suggestion : Il arrive parfois que l'utilisateur n'obtienne pas de résultats. En général, ce problème a lieu lorsque l'utilisateur a mal orthographié sa entrée. Le moteur de recherche Google a été le premier à lancer un correcteur orthographique pour suggérer des requêtes plus appropriées. A la rédaction du mémoire, la suggestion est également proposée sur les moteurs de recherche *Yahoo*[8] et *AltaVista*[9]. Étant donné que les noms de gènes ne sont pas toujours connus par coeur (voir section 3.1), la suggestion peut aider les biologistes à retrouver les gènes en utilisant la méthode illustrée dans la section 3.1.

Maintenant que nous avons défini la suggestion et l'estimation, nous présentons dans ce paragraphe leur implémentation. La suggestion et l'estimation ont la même façon de procéder :

1. Localiser le lieu d'arrêt de la recherche (c'est à dire le sommet ou l'arc où s'arrête l'algorithme d'Aho-Corasick pour la requête donnée).
2. Dénombrer le nombre de feuilles qui peuvent être atteint depuis le lieu d'arrêt. Ce dénombrement peut être fait soit par un algorithme d'exploration en largeur soit par un algorithme d'exploration en profondeur.

[8]http ://www.yahoo.com
[9]http ://www.altavista.com

L'algorithme d'estimation, à ce stade, a terminé sa tâche et il retourne à l'utilisateur le nombre de feuilles qui peuvent être atteintes[10] à partir du lieu d'arrêt (chaque feuille représente une entrée du dictionnaire selon notre définition).
L'algorithme de suggestion doit faire une étape additionnelle. Le but de la suggestion est de générer une liste d'appellations similaires à l'entrée. Chacune de ses appellations va être une combinaison d'une partie commune et une partie propre. La partie commune représente le chemin de la racine au lieu d'arrêt. La chaîne de caractères correspondante à ce chemin est présente dans chacune des appellations suggérées. Chaque appellation suggérée va avoir également une partie propre qui la différentie des autres. Il s'agit d'une chaîne de caractère qui représente un chemin entre le lieu d'arrête et une feuille donnée. La suggestion peut être considérée comme un cas spécial de *word completion*[11].

Dans le paragraphe précédent, nous avons parlé de l'implémentation de la suggestion et l'estimation. Dans ce paragraphe, nous discuterons des techniques qui ont été mises en place pour permettre un délai de réponse acceptable pour une plate-forme en ligne. Pratiquement, plus de 400 arborescences sont utilisées pour l'estimation et la suggestion (Tableau 3.5). Ce nombre élevé s'explique par :

- Souci de performance
- Sens de lecture
- Restriction sur l'espèce

Performance : Le génome humain contient à peu près 40,000 gènes et celui de la souris en contient à peu près 60,000. Il est évident que certaines mesures doivent être prises pour conserver un temps de réponse acceptable. En général, lorsque la taille d'un graphe devient élevée, il est conseillé de faire une construction partielle

[10] Étant donné que notre plate-forme est un site Web, nous avons décidé de retourner l'estimation du nombre de résultats au lieu des résultats eux-mêmes.

[11] *word completion* est un outil commun qui existe dans les fureteurs en général. Lorsque l'utilisateur commence à taper un mot fréquemment utilisé, l'ordinateur le complète ou propose une liste de choix.

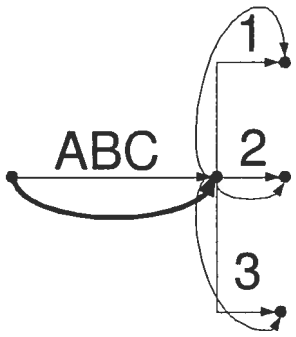

FIG. 3.8 – **Le processus de suggestion pour *"ABC"* :** Pour effectuer une suggestion, le moteur de recherche commence par lancer une recherche et trouver le lieu d'arrêt de la recherche. Ensuite, il va dénombrer toutes les feuilles qui peuvent être atteintes à partir de ce lieu d'arrêt. Chaque appellation suggérée est donc une combinaison d'une partie commune (flèche en gras) et d'une partie propre (flèches normales). La partie commune est le chemin de la racine au lieu d'arrêt et la partie propre est un chemin du lieu d'arrêt à une feuille donnée.

du graphe et de ne charger que ce qui est nécessaire. Nous avons opté pour cette mesure en divisant l'arborescence de l'algorithme d'Aho-Corasick en plusieurs arborescences. Le processus de division est assez simple. Chaque caractère[12] correspond à une arborescence. Lorsque l'utilisateur commence à taper sa requête, la plateforme décide quelle arborescence doit être chargée en mémoire. Ce choix dépend du sens de lecture (voir paragraphe suivant) choisi par l'utilisateur.

Sens de lecture : Nous avons discuté dans section 3.1 de la façon dont les utilisateurs se souvenaient des noms de gènes : *commençant par* ou *finissant par.* Le moteur de recherche permet ces deux types de recherche. La plupart des figures que nous avons montrées étaient du genre *commençant par.* Le concept *finissant par* est assez similaire à celui que nous avons illustré. La différence est que les mots

[12]lettre, chiffre ou symbole

sont inversés[13] avant d'être insérés dans l'arborescence correspondante qui est dite *inversée*. Par conséquent, il existe pour chaque caractère, une arborescence normale et une arborescence inversée. Dans cette dernière, la requête doit être inversée avant de lancer la recherche. Ainsi chaque caractère possède deux arborescences. La première contient toutes les appellations qui commencent par ce caractère. Et la deuxième va contenir toutes les appellations qui finissent par ce caractère.

Restriction sur l'espèce : Étant donné que le dictionnaire (dont nous discuterons plus tard) contient deux espèces à date : la souris et l'humain. L'utilisateur doit être en mesure de restreindre l'espace de recherche à l'une ou à l'autre des espèces. Par conséquent, il existe 3 types d'arborescences :

- celles contenant les appellations de gènes de la souris seulement
- celles contenant les appellations de gènes de l'humain seulement
- celles contenant les appellations de gènes de la souris et de l'humain

Nous avons tenté de regrouper les arborescences mais cette tentative s'est traduite par une augmentation significative du temps de réponse.

	Lettres	Chiffres	Autres	Total
Toutes les espèces	52 (52)	10 (10)	4 (14)	66 (76)
Homo Sapiens	52 (52)	10 (10)	2 (9)	64 (71)
Mus Musculus	52 (52)	10 (10)	4 (14)	66 (76)

TAB. 3.3 – **Nombre d'arborescences utilisées pour l'estimation et la suggestion :** Ce tableau montre le nombre d'arborescences utilisées par espèces et par type de caractères (c'est-à-dire le premier caractère du mot ou le dernier caractère du mot (Ce choix dépend du sens de lecture). Le moteur de recherche a deux sens de lecture : *commençant par* ou *finissant par*. Les nombres entre parenthèses représentent les arborescences servant à la recherche *finissant par*. Le nombre total des arborescences utilisées est obtenu en ajoutant les nombres sans parenthèses à ceux avec parenthèses dans la section total.

[13] l'inverse du mot ABC est CBA

3.6 Chromosome Browser

Les appellations de gènes et les identifiants de base de données sont très pratiques pour accéder à l'information des gènes lorsqu'ils sont connus de manière approximative ou exacte. Mais nous avons vu dans la section 3.1 que même les approximations pouvaient contenir des erreurs. De plus, dans la section 1.4, nous avons souligné que les auteurs n'utilisaient pas toujours les appellations de gènes qui permettent de trouver l'information dans une base de données. Dans cette section, nous présentons un moyen pour retrouver un gène à partir de son emplacement chromosomique.

3.6.1 Concept

En général, lorsqu'un utilisateur ne connaît pas le nom de l'objet recherché, il tente de le retrouver par son contenu ou sa description. Par exemple, le nom d'une chanson peut être trouvé en tapant les paroles dans un moteur de recherche tel que Google. Un livre ou un auteur peuvent aussi être retrouvés de la même manière en tapant un passage. La nouvelle génération de moteurs de recherche permettra peut-être de faire des recherches en soumettant une vidéo ou une photo. Pour les gènes, le contenu est la séquence qui malheureusement n'est pas toujours connue ou trop longue pour être recherchée ou mentionnée dans un article. L'élément qui peut jouer le substitut à la séquence est probablement l'emplacement chromosomique. De plus, ce dernier est souvent mentionné dans les articles (lorsqu'il est connu).

3.6.2 Fonctionnement

Le module *Chromosome Browser* permet de retrouver un gène à partir d'un emplacement chromosomique spécifié par l'utilisateur. Ce module permet trois types d'emplacement chromosomique. Deux de ces types sont spécifiques à la souris et le troisième est propre à l'humain. À la rédaction du mémoire, il n'y a pas de système uniforme pour l'emplacement chromosomique dans ces deux espèces.

Humain

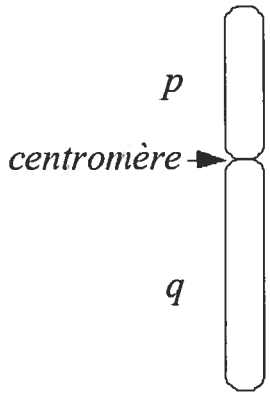

FIG. 3.9 – **Structure d'un chromosome Humain** : Chez l'humain, le chromosome est divisé en deux parties inégales par le centromère : les bras. Le bras long est dénoté par la lettre q et le bras court la lettre p. Pour localiser un gène humain, on utilise le chromosome, le bras et un entier. Ce dernier permet de localiser le gène précisément sur un bras donné.

Nous savons que les chromosomes contiennent les gènes. Les chromosomes se trouvent dans le noyau de la cellule. Une cellule humaine possède en général 23 paires de chromosomes. Dans chaque paire de chromosomes, un chromosome provient du père et un autre de la mère. Chacun des chromosomes a une forme caractéristique (Figure 3.9). Au centre du chromosome se trouve un étranglement de la structure appelé le *centromère*. Le centromère divise le chromosome en deux parties inégales. Ces deux parties sont appelées *bras*. Le petit bras est dénoté par la lettre p et le grand bras par la lettre q. Chez l'humain, un gène est localisé par le chromosome, le bras (petit ou grand) et un nombre réel. Le nombre réel représente une bande colorée qui apparaît lorsque le chromosome est décoloré avec des produits chimiques. Ce nombre mesure également la distance qui sépare l'emplacement du gène et le centromère. Plus cette distance est grande, plus le gène est éloigné du centromère. Ce type d'emplacement chromosomique est en général appelé emplacement *cytogénétique*. Par exemple, l'emplacement du gène p53 (TP53) est 17p13.1 (Figure 3.10), c'est à dire le petit bras du dix-septième chromosome.

Pour l'humain, Chromosome Browser n'oblige pas l'utilisateur à spécifier un emplacement chromosomique précis[14]. Il peut même spécifier un intervalle. De plus, les trois variables (chromosome, bras et réel) permettent différentes possibilités de recherche pour l'emplacement chromosomique humain (Tableau 3.4).

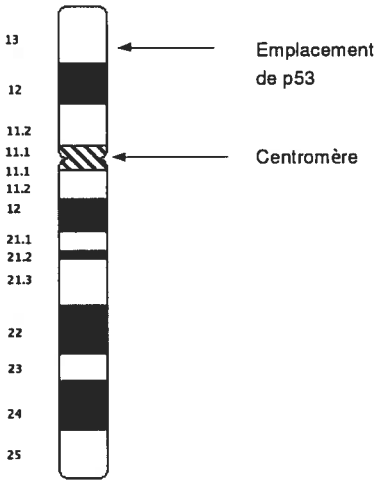

FIG. 3.10 – **Emplacement du gène p53 sur le chromosome 17 de l'humain** : Sous l'effet de produits chimiques, plusieurs bandes de couleurs différentes apparaissent sur le chromosome. Ces bandes permettent de localiser un gène sur un bras donné. Par exemple, le gène *TP53* est situé sur le petit bras (p) du 17ème chromosome ou précisément 17p13.1.

Souris

Chez la souris, il y a 20 paires de chromosomes. L'organisation structurale du chromosome est également différente. Les chromosomes de la souris n'ont pas de centromère qui se situe au milieu approximativement (comme chez l'humain). Leur centromère est situé sur le haut du chromosome [WLTB+02]. Il existe deux méthodes pour localiser un gène chez la souris : l'emplacement cytogénétique et

[14]Le fait que l'humain n'ait qu'un seul type d'emplacement chromosomique facilite la mise en place d'une certaine flexibilité pour ce mode de recherche. Dans la souris, il existe deux types d'emplacement. Nous n'avons pas trouvé une table de correspondance entre les deux types. Par conséquent, nous avons fait le choix de ne pas offrir de flexibilité dans la souris.

- chromosome seulement (c. à d. 1−22, X, Y, M).
- chromosome (c. à d 1−22, X, Y) et bras (cen, p, q).
- chromosome (c. à d 1−22, X, Y), bras (cen, p, q) et limite inférieure.
- chromosome (c. à d 1−22, X, Y), bras (cen, p, q) et limite supérieure.
- chromosome (c. à d 1−22, X, Y), bras inférieur (cen, p, q), limite supérieure et bras supérieur (cen, p, q).
- chromosome (c. à d 1−22, X, Y), bras inférieur (cen, p, q), bras supérieur (cen, p, q) et limite supérieure.
- chromosome (c. à d 1−22, X, Y), bras inférieur (cen, p, q), bras supérieur (cen, p, q).
- chromosome (c. à d 1−22, X, Y), bras (cen, p, q), limite inférieure et limite supérieure.
- chromosome (c. à d 1−22, X, Y), bras inférieur (cen, p, q), limite inférieure, bras supérieur (cen, p, q) et limite supérieure.

TAB. 3.4 – **Les différentes possibilités de recherche pour l'emplacement chromosomique humain :** Il existe plusieurs possibilités pour retrouver un gène humain en utilisant Chromosome Browser. Les chiffres 1 à 22 et les caractères X et Y servent à identifier les chromosomes. La lettre M est une abréviation de mitochondrie (qui fait partie de la cellule) où est stockée l'énergie nécessaire au fonctionnement de la cellule. Les mitochondries ont une propriété spéciale, ils contiennent certains gènes.

l'emplacement physique.

Emplacement cytogénétique Sous l'effet de produits chimiques, nous savons que des bandes de différentes couleurs apparaissent sur le chromosome. Chez la souris, les biologistes ont préféré identifier chaque bande par un caractère et un entier. Le gène de la souris *Uty* (*Entrez Gene* ID : 22290) a pour emplacement Y A1 c'est à dire le chromosome sexuel Y et la première bande grise (Figure 3.11). Avant d'introduire le deuxième type d'emplacement, nous allons introduire la méiose qui est un type de division cellulaire. Ce deuxième type d'emplacement utilise un phénomène qui se passe durant la méiose pour localiser un gène.

Méiose Il existe deux types de divisions cellulaires : La mitose et la méiose. La mitose permet de produire deux clones d'une cellule donnée. Si la cellule de départ

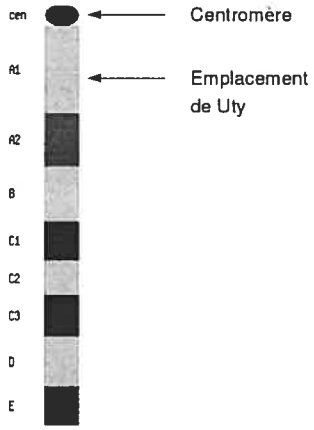

FIG. 3.11 – **Emplacement du gène Uty sur le chromosome Y de la souris** : Sous l'effet de produits chimiques, plusieurs bandes de couleurs différentes apparaissent sur le chromosome. Par exemple, le gène *Uty* se trouve sur la première bande colorée (juste en dessus du centromère).

possède 2n chromosomes, les cellules clones auront également 2n chromosomes. La méiose quant à elle, permet de produire 4 cellules filles ayant chacune la moitié des chromosomes de la cellule mère.
La mitose et la méiose sont tous deux précédées d'un dédoublement du matériel chromosomique. La mitose permet de renouveler les cellules d'un organisme donné tandis que la méiose permet de produire des gamètes (pour la reproduction asexuée). La méiose a une caractéristique intéressante. Durant le processus de la méiose, chaque paire de chromosomes peut échanger du matériel chromosomique. Ainsi, après la méiose, les deux chromosomes de chaque paire sont différents des deux chromosomes de départ. Maintenant que nous avons défini la méiose, nous pouvons introduire le deuxième type d'emplacement.

Emplacement physique Ce type d'emplacement se base sur les fréquences des échanges de matériel chromosomique durant la méiose. Une unité de mesure existe pour mesurer la probabilité d'échange sur une distance. C'est le *centiMorgam* Abréviation **cM**. Une unité de centiMorgan représente un crossing-over (échange de matériel chromosomique) sur 100 méioses. Plus cette distance est grande entre

deux gènes donnés, plus il y a de chances d'échanges de matériel durant la méiose. Par exemple, l'emplacement du gène *p53* est 11 39.0 cM et celui de *Egfr* est 11 9.0 cM, par conséquent il y a de grandes chances d'échange de matériel chromosomique entre les deux (Figure 3.12). Dans ce type d'emplacement, chaque chromosome est gradué en centiMorgan. 0 cM se trouve en haut du chromosome (coïncidant avec le centromère).

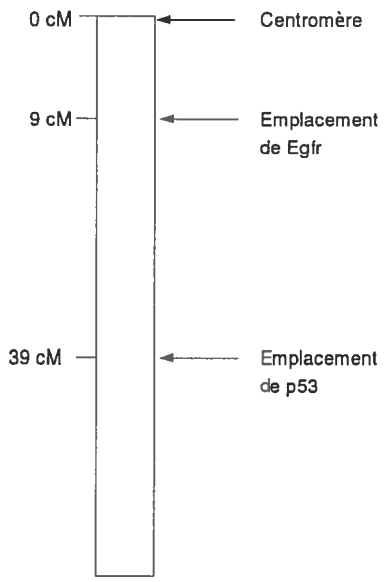

FIG. 3.12 – **Emplacement des gènes p53 et Egfr sur le chromosome 11 de la souris** : Le chromosome est gradué comme une règle du haut vers le bas. La graduation commence au centromère qui est situé sur le haut du chromosome. Le gène Egfr est plus proche du centromère que p53. L'unité centiMorgan représentant également une probabilité d'échange de matériel durant la méiose, les gènes p53 et Egfr ont de grandes chances d'échanger du matériel génétique à cause de la grande distance les séparant.

Le processus de recherche avec Chromosome Browser est simple. L'utilisateur soumet un emplacement chromosomique (en respectant l'un des trois types d'emplacement énoncés précédemment). Le module retourne une liste de gènes et l'utilisateur a le choix soit de consulter chacun d'entres eux ou d'utiliser les informations disponibles (nom et symbole officiels... etc) sur la page de résultats pour retrouver

– chromosome seulement (c. à d. 1−19, X, Y, M).
– chromosome (c. à d 1−22, X, Y) , bande cytogénétique (c. à d. A, B, C, D, E, F, G, H) et un nombre entier (c. à d 1, 2, 3, 4, 5, 6).
– chromosome (c. à d 1−22, X, Y) et distance en centiMorgan (c. à d. 0.00 − 115.80 **cM**)

TAB. 3.5 – **Les différentes possibilités de recherche pour l'emplacement chromosomique de la souris :** Il existe plusieurs possibilités pour retrouver un gène de la souris en utilisant Chromosome Browser. Les chiffres 1 à 19 et les caractères X et Y servent à identifier les chromosomes. La lettre M est une abréviation de mitochondrie (qui fait partie de la cellule) où est stockée l'énergie nécessaire au fonctionnement de la cellule. Les mitochondries ont une propriété spéciale, ils contiennent certains gènes.

son gène.

Une fois la construction de l'outil terminée, nous avons voulu déterminer la performance de l'outil et si possible le comparer avec des outils similaires. Ces deux analyses sont effectuées dans la section suivante.

3.6.3 Comparaison

Avant la construction de l'outil Chromosome Browser, il n'existait aucun outil qui permettait de retrouver un gène à partir de son emplacement chromosomique. L'absence de ce type d'outil a été la raison principale au développement de Chromosome Browser. Nous avons été également motivés par l'utilisation des appellations non-officielles par les biologistes dans leurs articles (voir section 1.4). A la fin de la construction de l'outil, la situation de départ avait changé. Les moteurs de recherche Entrez Gene et HGNC proposent maintenant des outils similaires. Entrez Gene propose maintenant de rajouter à la requête le chromosome mais il ne permet pas un emplacement chromosomique précis. Un chromosome peut contenir jusqu'à un millier d'emplacements chromosomiques par conséquent la requête doit être combinée avec des mots clefs pour éviter des milliers de résultats. HGNC propose un moyen plus intéressant. L'utilisateur peut spécifier un emplacement chromosomique qui {commence par, finit par, contient, est égal à} la requête. HGNC est sans

aucun doute une alternative fiable et efficace à Chromosome Browser. Maintenant que ce dernier a un concurrent, nous avons voulu comparer son utilisation et son efficacité. Nous avons extrait 3 exemples de la littérature où un tel type d'outil est nécessaire.

Un emplacement approximatif La littérature des années 80 a été florissante dans le domaine biomédical tout particulièrement pour la détermination de l'emplacement des gènes connus à l'époque. Les techniques des années 80 n'étaient pas aussi précises qu'elles ne le sont aujourd'hui et les biologistes utilisaient alors des approximations. Il faut également souligner que les démarches de formalisation étaient à leur début donc il se peut très bien que le nom du gène mentionné ait changé. Dans un des articles de l'époque, nous avons trouvé la phrase suivante :
"LARS gene (previously referred to as leuS), which is located near the centromere on 5q ..." [MCWA85]
Nous savons que le système d'emplacement chromosomique chez l'humain repose sur le chromosome, le bras et un nombre réel qui représente la distance séparant le gène du centromère. Plus le gène est proche du centromère, plus cette distance est petite. Le gène *LARS* est selon l'article proche du centromère par conséquent la distance le séparant du centromère va être petite. Nous pouvons même donner un intervalle de valeurs possibles disons 5q01-q20 ou [01,20] sur le grand bras du chromosome 5. Cette mise en intervalle se prête très bien à Chromosome Browser. Pour HGNC, deux requêtes doivent être effectuées c'est-à-dire rechercher tous les emplacements commençant par *5q0* puis ceux commençant par *5q1*. Le nombre de résultats diffère pour les deux outils et la raison en est simple. Chromosome Browser recherche dans tous les gènes connus actuellement tandis que HGNC recherche que dans ceux qu'il a approuvés (voir section 1.5). Finalement, les deux outils font match nul et *LARS* n'est pas trouvé à partir de son emplacement. Nous découvrirons dans la littérature que le gène ne se trouvait finalement pas proche du centromère. Une rectification de son emplacement à ce propos a été effectuée récemment. Ce changement d'emplacement pose en général des problèmes pour les

bases de données. Nous discuterons de ce problème dans le chapitre suivant.

Une région approximative Souvent les biologistes vont s'intéresser à des régions du chromosome et encourager des chercheurs à lancer des études complémentaires. C'est le cas du Docteur Stoll qui mentionne le cas d'un patient qui pourrait être un exemple humain de progenèse[15] [SMP05]. Dans son étude, la docteure a remarqué que son patient avait une région chromosomique manquante. La région *13q21-q31* avait été supprimée du génome de l'individu. Dans sa conclusion, elle fait le lien entre la région supprimée et la maturité sexuelle précoce de l'individu. Pour elle, il ne fait pas de doute que des gènes dans cette région contrôlent le développement sexuel. Nous avons voulu extraire les gènes connus dans cette région à l'aide des deux outils. Une simple manipulation avec Chromosome Browser a permis d'extraire 102 gènes tandis deux manipulations simples avec HGNC ont permis d'extraire 44 gènes. Cette différence du nombre de résultats s'explique encore une fois par la portion des gènes qui ne sont pas encore approuvés par HGNC.

Une région précise Dans le paragraphe précédent, nous avons donné l'exemple d'une région approximative. Mais il existe également des articles qui mentionnent des régions précises. Le trouble bipolaire ou la maniaco-dépression est un trouble qui est caractérisé par des périodes d'hyperactivité et de bonheur suivi par des périodes de tristesse et de dépression. Ce trouble est assez répandu et par conséquent, un certain nombre d'articles sont faits sur le sujet. Dans son article [FLSG+06], Docteur Steven Faraone suspecte qu'un gène lié au trouble bipolaire se trouve dans la région *9q34*. Encore une fois, le chercheur encourage des recherches dans cette région. Nous avons encore une fois tenté d'extraire les gènes présents dans cette région à l'aide de ces deux outils. 293 gènes et 229 gènes ont été respectivement trouvés par Chromosome Browser et HGNC.

Discussion : Nous aurions voulu comparer Chromosome Browser avec MGI mais

[15] La progenèse est une atteinte dans l'age juvenile de la maturité sexuelle.

malheureusement ce dernier ne propose pas encore un tel outil. Dans ce cas, nous discuterons que de la comparaison HGNC/Chromosome Browser. Dans notre comparaison, Chromosome Browser semble avoir un certain avantage sur HGNC. Cependant, cet avantage est temporaire car le but de cet organisme est de couvrir tous les gènes humains. Pour le génome humain, nous avons quand même trouvé que HGNC demandait un certain travail supplémentaire car sa flexibilité n'est pas adaptée aux chiffres. Un chiffre en général, est identifié par un intervalle et non pas par son contenu. Le chiffre 12 à titre d'exemple se trouve entre 11 et 13. Cette identification est plus forte que de mentionner les chiffres qui composent 12 (car 21 partage strictement les mêmes chiffres et 120 le même début).

Cependant, l'apparition de ces outils marque un tournant dans la recherche des gènes. Maintenant, un gène peut être retrouvé à partir de son emplacement. Dans notre exemple, nous avons également montré qu'il existait des applications pour ce genre d'outil. Au début, nous pensions que cet outil pouvait être une alternative pour rechercher un gène précis quand le moteur de recherche principal échoue. Finalement, nous réalisons que cet outil simple peut servir dans l'extraction des gènes présents dans une région chromosomique donnée. Nos exemples ont montré que cette fonction était même plus utile que la fonction originelle de Chromosome Browser.

Dans l'introduction, nous avons souvent critiqué les moteurs de recherches qui retournaient beaucoup de résultats. Dans la section suivante, nous étudions le nombre de résultats qui est généré par Chromosome Browser.

3.6.4 Étude de densité des chromosomes

Suite à nos critiques parfois sévères contre certains moteurs de recherche, nous avons voulu déterminer si Chromosome Browser était un outil qui générait beaucoup de résultats. Pour cela, nous avons :

1. Téléchargé tous les emplacements chromosomique de la base de données[16]

[16]La base de données sera présentée dans le chapitre suivant.

2. Éliminé de la liste toutes les redondances
3. Pour chaque emplacement chromosomique de la liste, mesuré le nombre de gènes correspondants
4. Généré des statistiques

Espèce	Humain	Souris
minimum	1	1
maximum	765	1302
Nombre total de gènes	43468	57305
Nombre d'emplacements	3140	4759
Nombre Moyen de gènes par emplacement chromosomique	13.84	12.04
Écart Type	36.86	62.85

TAB. 3.6 – **Résultats de l'analyse des emplacements chromosomiques trouvés dans la base de données :** Nous avons extrait tous les emplacements possibles de gènes dans notre base de données (Voir chapitre suivant). Nous avons éliminé la redondance dans cette liste et nous avons mesuré pour chaque emplacement le nombre de gènes correspondant.

Les résultats de ce test sont présentés au tableau 3.6. Nous avons trouvé respectivement 3140 et 4759 emplacements chez l'humain et la souris. A première vue, les résultats semblent être satisfaisants. En nous basant seulement sur la moyenne, une recherche de gènes à partir d'un emplacement chromosomique donné va générer en moyenne entre 12 et 14 résultats. Malheureusement, ce bon résultat n'est pas tout à fait véridique. Il suffit de s'attarder un peu sur l'écart type pour remarquer un fait important. L'écart type des deux organismes est assez élevé par rapport à la moyenne. L'écart type chez l'humain fait trois fois à peu près la moyenne tandis que chez la souris c'est à peu près 6 fois la moyenne correspondante. En général, lorsque l'écart type est grand, alors il existe de grandes différences entre les valeurs. Nous avons eu du mal à comprendre des écarts types aussi importants. Nous avons alors décidé de calculer la médiane. La médiane divise un ensemble de valeurs entre deux ensembles égaux. Dans la première partie, ce sont tous les chiffres qui sont

inférieurs ou égaux à la médiane. Dans l'autre partie, ce sont les chiffres supérieurs à la médiane. À notre grande surprise, la médiane s'est avérée être une petite valeur chez les deux espèces : 2. Cette petite médiane explique donc la grandeur de l'écart type. L'écart type est une mesure moyenne de l'écart entre les différentes valeurs et la moyenne. La médiane étant petite, la moyenne va être forcement entraînée vers une petite valeur. Cependant, ce sont les grandes valeurs qui vont faire en sorte que l'écart type soit important.

Nous avons voulu illustrer cette remarque par un graphique qui mesure le nombre d'emplacements correspondant à un certain nombre de gènes (humain : figure 3.13/ souris : figure 3.14). Malheureusement, nous avons fait le choix de montrer qu'une partie du graphique. La raison en est simple. Les emplacements qui génèrent peu de résultats sont très nombreux et par conséquent, ils cachent les autres emplacements. Les deux figures illustrent que les emplacements correspondants à de nombreux gènes diminuent en même temps que le nombre de gènes correspondants augmente. Cependant, il n'y a pas à proprement dit une concentration à un endroit précis de la courbe. La valeur maximale de gènes par emplacement est atteinte à 1302 et 765 respectivement chez la souris et chez l'humain. Nous remarquons également que la courbe a la forme de la fonction inverse $f(x)=1/x$.

Avant d'étudier toutes les implications que ces chiffres pouvaient avoir sur la recherche, nous avons voulu lancer un dernier test. Nous avons voulu tout particulièrement comprendre pourquoi certains emplacements généraient beaucoup de résultats et d'autres un nombre minime. Nous voulions savoir si la précision de l'emplacement avait un rapport avec le nombre de résultats. Nous savons que certains gènes, à date, ne sont connus précisément c'est-à-dire que nous connaissons leur chromosome mais pas leur emplacement précis sur ce dernier. Or un chromosome peut contenir des milliers d'emplacements et peut expliquer le nombre de résultats excessifs. Pour répondre à notre question nous avons décidé d'analyser les emplacements chromosomiques selon leur position par rapport à la médiane et de les classer dans deux catégories chez la souris et trois catégories chez l'humain. Les trois catégories d'emplacements chez l'humain sont

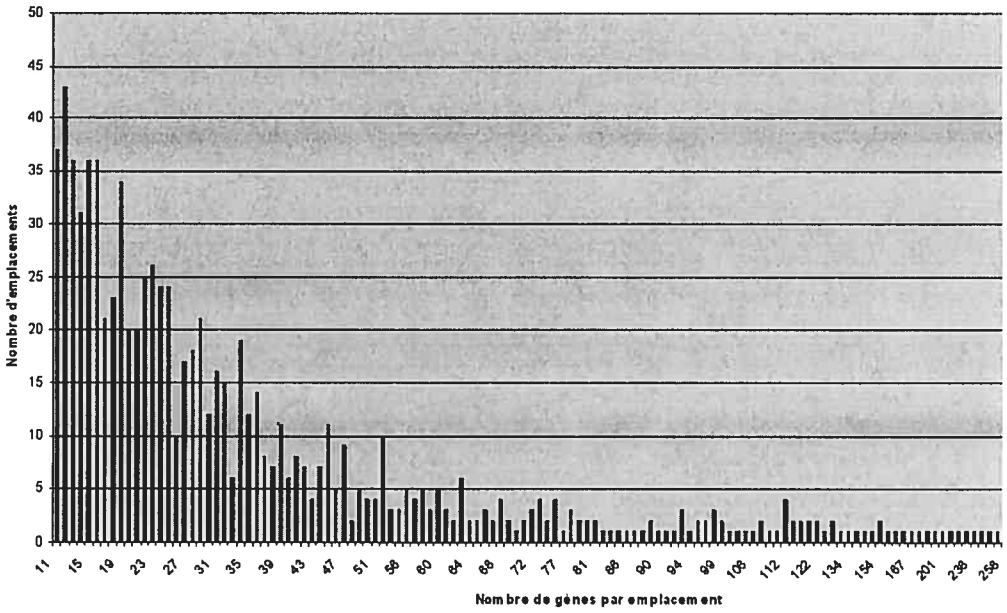

FIG. 3.13 – **Nombre de gènes par emplacements tirés de la base de données chez l'humain.** L'histogramme représente le nombre d'emplacements qui correspondent entre 11 et 258 gènes. Nous remarquons que plus le nombre de gènes augmente plus le nombre d'emplacements correspondants diminue.

1. Emplacement précis (le chromosome, le bras et la distance)
2. Emplacement à moitié précis (le chromosome et le bras seulement)
3. Emplacement flou (le chromosome seulement)

Et les deux catégories d'emplacements chez la souris sont :

1. Emplacement précis (le chromosome et soit une lettre soit une distance en centiMorgan)
2. Emplacement flou (le chromosome seulement)

Finalement les résultats ont prouvé que la précision chez l'humain et la souris n'avait aucun rapport avec le nombre de résultats. Nous avons trouvé à notre grande surprise que les emplacements précis dominent les autres emplacements quelque soient leurs positions par rapport à la médiane. Nous espérions que les

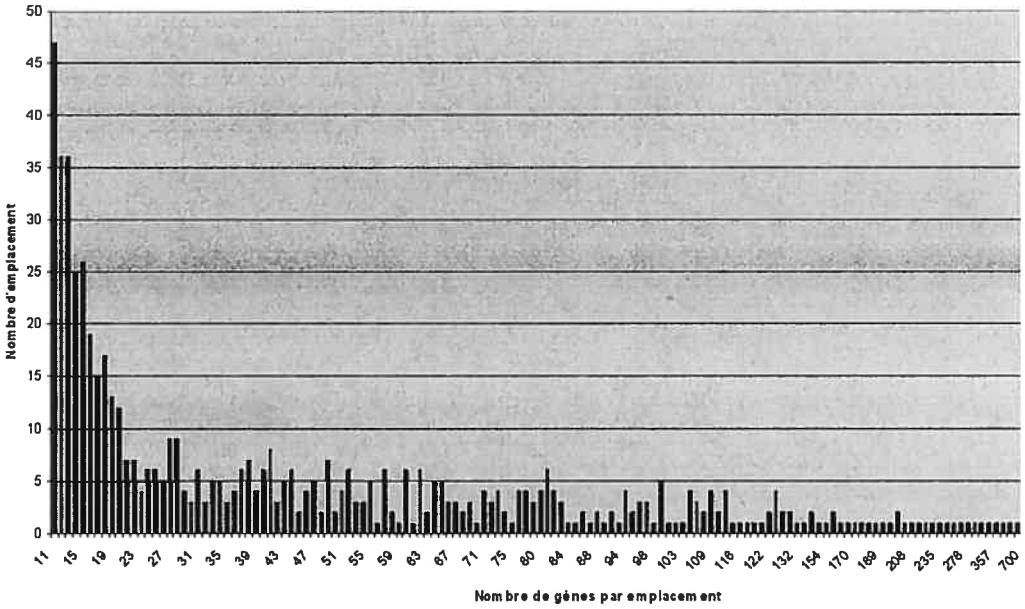

FIG. 3.14 – **Nombre de gènes par emplacements tirés de la base de données chez la souris.** L'histogramme représente le nombre d'emplacements qui correspondent entre 11 et 700 gènes. Nous remarquons que plus le nombre de gènes augmente plus le nombre d'emplacements correspondants diminue.

emplacements situés après la médiane tombent dans la catégorie des flous mais ça n'a pas été le cas.

Discussion Avant de nous lancer dans la discussion, il faut souligner que nous avons basé nos deux études sur les emplacements présents dans notre base de données. Par conséquent, une certaine précaution doit être prise. Ces emplacements représentent une portion des emplacements possibles. Il est difficile de définir statistiquement le pourcentage précis de cette portion. De plus, nous n'avons pas trouvé la même structure pour les emplacements de gènes. Certains emplacements de gènes se trouvent dans une bande tandis que d'autres s'étendent sur deux bandes. Cette remarque est particulièrement pertinente pour les gènes humains.
Néanmoins, étant donné que les emplacements représentent une portion réelle des emplacements possibles, une discussion est envisageable sur le sujet. Les deux

	Précision	Humain	Souris
Avant la médiane	précis	99%	100%
	semi-précis	0.60%	–
	flou	0.40%	0%
Après la médiane	précis	96.88%	99.19%
	semi-précis	1.74%	–
	flou	1.38%	0.81%

TAB. 3.7 – **Résultats de l'analyse de la précision des emplacements chromosomiques selon leur position par rapport à la médiane.** Dans notre analyse, un emplacement précis chez l'humain contient le chromosome, le bras et la distance. Un emplacement semi-précis contient le chromosome et le bras seulement. Chez la souris, un emplacement précis contient la chromosome et soit la distance en centiMorgan soit l'identifiant de la bande colorée. A partir de cette analyse, nous remarquons que la précision de l'emplacement a peu d'influence sur le nombre de gènes correspondants.

études prouvent que Chromosome Browser est un outil qui va soit générer beaucoup de résultats soit très peu de résultats. Cette remarque est également valable pour HGNC. Il ne semble pas avoir un juste milieu. L'utilisateur risque soit de tomber sur un gène, soit directement sur une trentaine. La dernière étude démontre également que la précision ne va pas forcément aider l'utilisateur. En règle général, rajouter des mots clefs ou être plus précis permet de réduire les résultats. Ici, l'effet inverse risque de se produire. Nous avons songé aux moyens de réduire l'ampleur de ce problème. Malheureusement, le nombre de gènes par emplacement est intimement lié aux progrès et aux découvertes en biologie. Nous pensons que ce nombre va probablement réduire dans le futur avec l'arrivée de technologies plus avancées et plus précises. En attendant, le fait de générer un grand nombre de résultats reste acceptable. De plus, une différence claire existe avec les outils que nous avons critiqués. Ces derniers retournent un ensemble de gènes liés directement ou indirectement à l'entrée recherchée. Certaines recherches peuvent avoir besoin de ce type d'information mais pas toutes. Le parallélisme doit être plutôt tracé avec les homonymes. Un homonyme correspond à plusieurs gènes. Il est donc normal qu'un outil retourne

un ensemble de gènes pour un homonyme donné. Ces gènes sont reliés directement car ils partagent une appellation en commun. La même remarque est valable pour les gènes partageant un même emplacement. Cependant, certaines options doivent être mises en place pour faciliter la recherche à l'aide de Chromosome Browser. Nous discuterons de ces options dans le chapitre 6.

CHAPITRE 4

LE DICTIONNAIRE

Dans le chapitre précédent, nous avons discuté de deux outils. Le premier outil est destiné à retrouver un gène à partir d'une de ses appellations ou de l'approximation de celle-ci. Cet outil utilise également un estimateur du nombre de résultats pour réduire les résultats excessifs et un outil de suggestion pour guider l'utilisateur lorsque sa requête échoue. Le deuxième outil permet également de retrouver un gène mais cette fois à partir de son emplacement sur le chromosome. Dans ce chapitre, nous présentons la source d'information du moteur de recherche : le dictionnaire. Normalement, le dictionnaire est un détail de l'implémentation. En biologie, une mauvaise organisation de la base de données et de sa mise à jour peut avoir des effets néfastes [SMWK06].

4.1 Le contexte biologique

La biologie est une science où les connaissances sont souvent remises en cause par des découvertes. Les gènes ne sont pas une exception à cette règle. Nous avons déjà mentionné les miARN [ABB+03] qui ont bouleversé notre compréhension du fonctionnement de la cellule et surtout des mécanismes de régulations des gènes. Pour les gènes, les difficultés se posent en général pour l'emplacement chromosomique (Voir section 3.6.3) et la fonction. La fonction principale d'un gène est de porter l'information génétique mais on lui assigne en général une autre fonction qui est reliée au rôle de sa protéine. Si la protéine correspondante est responsable de la respiration de la cellule, on dit que la fonction du gène est la respiration. Un gène est souvent référé par sa fonction et il arrive que le nom d'un gène change à la découverte d'une nouvelle fonction ou de la négation de sa fonction originelle. Le gène BCRA2 (*Breast Cancer 2, Early Set* ; Entrez Gene ID : 675) par exemple s'appelait FAS1 (*Fanconi Anemia Complementary Group D1*). Ce changement pose

des problèmes lors de la recherche dans la littérature ou sur Internet. L'utilisateur qui recherche BCRA2 risque de manquer certaines informations sur ce gène à commencer par son ancien nom et la littérature correspondante. Le résultat d'une telle recherche dépend surtout de la base de données utilisée et de la politique de mise à jour de cette dernière.

L'emplacement chromosomique d'un gène est également sujet à discorde. Le gène ME2 (*malic enzyme 2, NAD(+)-dependent, mitochondrial* ; Entrez Gene ID : 4200) possède deux emplacements suspectés : 6p25-p24 et 18q21. Par conséquent, l'architecture de la base de données doit être flexible pour supporter de tels changements et en même temps stricte pour maintenir la cohérence de l'information.

En biologie, les changements et les nouvelles informations sont fréquents. Dans Entrez Gene, il y a eu respectivement 10,639 et 29,113 créations ou suppressions d'information de gènes entre Janvier 2004 et Mars 2007 respectivement chez l'humain et la souris. Ces changements sont équivalents à respectivement 6 et 29 modifications majeures par jour chez l'humain et la souris. Ces chiffres malheureusement, n'englobent pas toutes les modifications mineures sur l'information d'un gène donné. Dans la section suivante, nous allons présenter la structure de notre base de données et les moyens mis en place dans cette dernière pour permettre une certaine flexibilité.

4.2 La base de données

La base de données utilisée par le dictionnaire est extraite de la base données Entrez Gene [MOPT05] du National Center for Biotechnology Information (NCBI[1]). L'information extraite est divisée en plusieurs catégories : principal (main), emplacement chromosomique (location), alias du nom de gènes (genes_alias), alias du nom de la protéine (proteins_alias) et finalement séquences (sequence) (Figure 4.1). Chaque catégorie représente une table dans la base de données. L'organisation de la table a été modifiée plusieurs fois lors de son développement et l'entrée de

[1]http ://www.ncbi.nlm.nih.gov/About/index.html

nouvelles données. Cela a révélé certains problèmes que nous suspections. Nous en avons également découvert de nouveaux. Certains d'entre eux ont entraîné la modification d'un champ de la table (ex : noms de gènes dépassant les 200 caractères) et d'autres ont entraîné la modification de l'ensemble de la table.

Par exemple, nous ne pensions pas qu'un gène pouvait être connu sans savoir son emplacement chromosomique. Nous ne savions pas non plus qu'un gène donné puisse avoir plusieurs emplacements suspectés. Nous avons fait face à ce problème surtout dans l'espèce humaine où nous nous sommes retrouvés avec plusieurs emplacements pour un gène donné. Nous avons examiné ces emplacements et nous avons trouvé que certains étaient équivalents. L'emplacement *12p* est équivalent à l'emplacement *12p15*. La différence réside dans le degré de précision. Pour d'autres gènes, les emplacements n'étaient pas vraiment équivalents et il a fallu changer la structure de la table *location* pour permettre qu'un gène donné puisse avoir plusieurs emplacements. Malheureusement, ces emplacements risquent de changer tout comme la fonction d'un gène. Par conséquent, un outil doit être mis en place pour mettre de l'ordre dans la base de données. Cet outil est la mise à jour, nous la présentons plus tard dans ce chapitre.

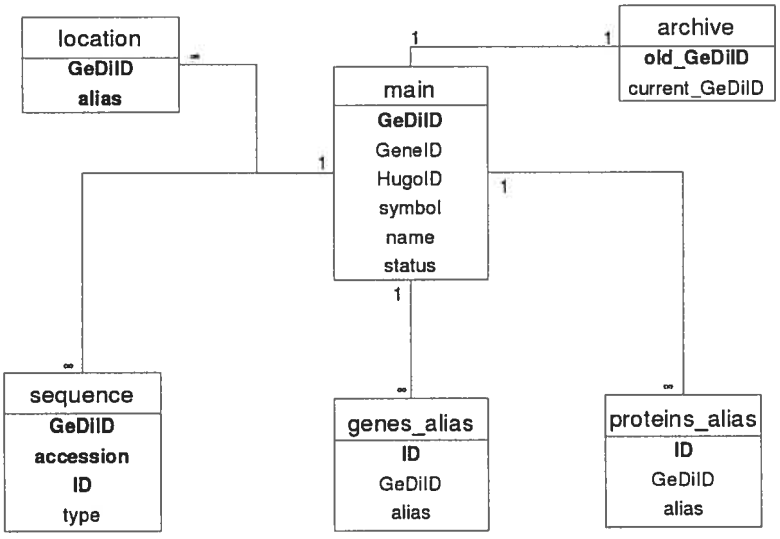

FIG. 4.1 – **Structure de la base de données**

Il est vrai que la source d'information du dictionnaire est la base de données Entrez Gene mais cette relation de dépendance est contrecarrée par certaines mesures qui ont été implémentées dans la plate-forme. Nous présentons également ces mesures dans les sections qui suivent. Une des première mesures que nous allons présenter est l'utilisation de l'identifiant de base de données comme substitut/complément à l'appellation dans les articles. Cependant avant d'introduire cette section, nous présenterons de l'extraction de l'information de la base de données Entrez Gene du site NCBI.

4.3 Extraction de l'information de Entrez Gene

Entrez Gene est une des rares bases de données qui propose le téléchargement complet de son contenu. MGI à titre d'exemple ne le permet pas et son contenu nous provient à travers Entrez Gene. Néanmoins, l'extraction de l'information dans Entrez Gene en soi n'est pas une chose facile. Entrez Gene représente une des base de données du NCBI. NCBI possède au total 24 autres bases de données. Il devient évident que les ingénieurs et administrateurs du NCBI placent chacun de leurs efforts sur toutes ces bases et non pas sur une en particulier. C'est un grand désavantage pour l'utilisateur. Ce dernier se retrouve avec des outils compatibles avec toutes les bases de données. Or la base de données des protéines ne contient pas les mêmes informations que celle des gènes. Le même argument s'applique aux fichiers à télécharger. Avant de discuter de ces derniers, nous allons introduire le langage de stockage de l'information du NCBI. Pour stocker les information, NCBI utilise l'*ASN*[2] pour les données telles que les séquences de nucléotides ou séquences d'acides aminés, les structures, les génomes ou encore les entrées MEDLINE (voir section 1.3). Ce format permet aux ordinateurs et programmes de tout type d'échanger d'une manière efficace les données. Un exemple d'une partie de l'information d'un gène encodé en ASN.1, est illustré à la figure 4.2.

[2]ASN.1, ou *Abstract Syntax Notation One*, est un format de représentation de données utilisé pour faciliter les échanges entre les plates-formes.

```
Entrezgene : := {
track-info {
geneid 7157,
status live,
create-date std {
year 1998,
month 8,
day 13
}, update-date std {
year 2007,
month 3,
day 11
}
}
```

FIG. 4.2 – **Exemple de l'information d'un gène en ASN :** Le format ASN peut être considéré comme une représentation d'un objet contenant plusieurs variables ou objets. Ici, l'objet *Entrezgene* contient deux objets *track-info* et *update-date std.* Cet exemple ne contient qu'une partie de l'information d'un gène.

Jusqu'au 8 février 2005, NCBI permettait de télécharger l'ensemble de l'information des gènes en ASN. Depuis lors, certaines modifications ont été effectuées. Le format des fichiers a changé et il a été remplacé par l'ASN binaire. Ce format à la différence de l'ASN n'est pas lisible par l'humain et doit passer par la machine. Cependant, NCBI propose des outils en C++ pour pouvoir extraire l'information. Malheureusement, ces outils ne sont pas adaptés à la simple extraction d'information. Dans la littérature, ces outils sont même considérés comme ayant une "courbe raide d'apprentissage" [LG05]. Nous avons alors tenté de rechercher un moyen plus simple d'extraire l'information. Nous avons finalement trouvé qu'il existait un moyen de convertir l'ASN binaire en XML en utilisant un outil du NCBI. Le format XML nous intéresse particulièrement car il permet une vérification des données extraites du fichier. Cette tâche est particulièrement importante lors des premières insertions. Malheureusement, notre prise en charge de cet outil n'a pas

été aussi directe. L'outil est conçu sans un mode d'aide pour informer l'utilisateur de la syntaxe appropriée. Nous avons tenté certaines manipulations qui ont été infructueuses avant de nous lancer dans une recherche exhaustive dans le site du NCBI. Cette recherche a été également infructueuse. Nous nous sommes alors tournés vers le moteur de recherche Google qui nous a permis de trouver la syntaxe appropriée sur un site indépendant du NCBI.

Le format XML du NCBI est assez ambigu car l'information peut paraître sous différentes balises et parfois sous un différent enchaînement de balises. Par exemple, pour extraire les deux identifiants de la séquence de l'ARN, il existe deux types d'enchaînements :

1. <Gene-commentary_type value="mRNA" >
2. <Gene-commentary_accession>identifiant1</Gene-commentary_accession >
3. <Object-id_id> identifiant2 <Object-id_id>

ou :

1. <Gene-commentary_type value="mRNA">
2. <Gene-commentary_accession>identifiant1</Gene-commentary_accession >
3. <Gene-commentary_Seq>
4. <Seq-id_gi>identifiant2</Seq-id_gi>

Le *DTD*[3] semble avoir été conçu pour plusieurs types d'information en biologie. Certaines balises sont très ambiguës (Par exemple : <Object-id_id>). Nous avons donc renoncé à utiliser le *DTD* et préféré faire une analyse manuelle des fichiers XML. Par conséquent, nous avons effectué à plusieurs reprises la comparaison entre l'information d'un gène en XML et en version texte pour déterminer les différents enchaînements possibles pour chaque information du gène.

Maintenant que nous avons discuté de l'extraction de l'information, nous pouvons parler de l'identification de l'information dans notre base de données. Ici, l'identification est différente dans sa structure mais également dans son utilisation.

[3]Document Type Definition est un document qui permet de décrire un modèle de document XML.

4.4 Identification des gènes dans la BD

Chaque base de données doit reposer sur un système d'identifiants. Pour les gènes, les bases de données utilisent en général des chiffres mais ce moyen efficace n'est pas populaire en pratique [TV06]. Ces identifiants sont ignorés par les biologistes. Ces derniers préfèrent utiliser des appellations dans leurs publications plutôt qu'un chiffre sans aucune signification [TV06] [WSVM+03]. Ce comportement comme nous l'avons vu, rend malheureusement inefficaces les techniques d'automatisation à cause de la confusion avec la langue anglaise et avec les termes biomédicaux (voir section 1.4). Les chiffres quant à eux, permettent de retracer facilement l'information dans une base de données et de faciliter ainsi la continuation de la recherche. Certains biologistes ont proposé des solutions. La solution la plus populaire est de mentionner dans l'article l'identifiant du gène dans une base de données [LH05]. À la rédaction du mémoire, les biologistes ne semblent pas avoir adopté cette solution.

Ainsi, nous avons réfléchi à un autre moyen. La solution serait de créer un identifiant significatif pour les biologistes mais en même temps de faciliter l'application d'outils d'extraction d'information. Il serait également intéressant de pouvoir facilement identifier des gènes d'un organisme spécifique pour éviter le temps en général perdu à le déterminer. Nous avons alors recherché un identifiant qui combine l'organisme et d'autres caractéristiques du gène. Cette grammaire devait pouvoir être généralisée à chacun des gènes d'un organisme donné ainsi qu'aux autres organismes.

Les identifiants de la base de données sont ainsi formés de trois parties (Figure 4.1). La première partie est composée de 5 caractères qui identifient l'organisme (HoSap pour l'humain et MuMus pour la souris). Cette première partie est créée en considérant le nom binomial[4] (*Homo Sapiens* et *Mus Musculus*) et en combinant les deux premiers caractères du nom générique ("Homo" et "Mus") aux trois pre-

[4]En taxinomie (botanique, zoologie, etc.), le nom binominal, ou binôme, est la combinaison de deux noms, servant à désigner une espèce.

miers caractères du nom spécifique ("Sapiens et Musculus"). Cette combinaison en respectant les majuscules, permet de nommer jusqu'à 11 millions d'espèces (26^5 : 26 étant le nombre de lettres dans l'alphabet) et jusqu'à 300 millions d'espèces si on permet une certaine flexibilité (52^5 : 26 + 26 étant le nombre de lettres en majuscule et en minuscule). Les estimations du nombre d'espèces sur terre varient entre 10 millions et 65 millions [Hun93]. La deuxième partie de l'identifiant doit pouvoir différencier le gène parmi les autres, et cette grammaire doit être en même temps généralisable à tous les gènes de cette espèce. Or chaque gène possède un emplacement chromosomique dans la souris et dans l'humain. Il suffit d'incorporer l'emplacement chromosomique dans l'identifiant (chromosome et bras pour l'humain et chromosome pour la souris). Finalement, la troisième partie doit pouvoir différencier les gènes de la même espèce qui partagent le même emplacement. Un chiffre est utilisé pour cette tâche.

Homo Sapiens	"HoSap"(i('p','q','cen','pq','U')‖'U')_z
Mus Musculus	"MuMus"(j‖'U')_z
où	
i ∈ {[1,22],'X','Y','M'}	
j ∈ {[1,19],'X','Y','M'}	
z ∈ ℕ*	

TAB. 4.1 – **Grammaire de l'identifiant pour l'humain et la souris :** chaque identifiant de gène est composé de trois parties. La première partie est dérivée du nom binomial et elle permet de définir l'appartenance d'un gène à une espèce donnée. La deuxième partie est dérivée de l'emplacement chromosomique et sert à différencier les gènes appartenant à la même espèce. Et la troisième partie permet de différencier les gènes appartenant au même chromosome.

Nous avons vu dans la section 4.1 qu'il pouvait exister un certain flou concernant l'emplacement chromosomique. L'identifiant doit traduire cette incertitude. C'est pour cela que la deuxième partie de l'identifiant peut contenir la lettre 'U'

pour *unknown*[5]. La lettre U dans ce contexte est utilisée pour les gènes sans emplacements connus mais également ceux avec plusieurs emplacements suspectés. Étant donné que cette incertitude est temporaire, la base de données doit détecter si l'emplacement a été déterminé dans le fichier de Entrez Gene. C'est le rôle de la procédure de la mise à jour que nous présentons dans la section suivante.

4.5 Mise à jour et cohérence

Avec la constance des découvertes, les mises à jour sont essentielles pour un dictionnaire de gènes [SMWK06]. La rapidité de l'émergence de nouvelles informations nécessite des nouvelles stratégies pour les bases de données publiques. Certains administrateurs vont faire des insertions de masses (introduisant des redondances possibles à court terme et des informations dispersées à long terme) et d'autres vont écraser les données existantes (perte possible d'informations). Nous avons développé plusieurs techniques pour éviter certains de ces problèmes (Tableau 4.2). Lors d'une mise à jour, les nouvelles données n'écrasent pas les anciennes. Par exemple lors du changement du nom et symbole officiel, les anciennes informations vont directement dans les alias du gène. Nous avons également mentionné précédemment que l'identifiant du gène dépend de son emplacement chromosomique. Lors du changement de l'emplacement chromosomique ou de sa détermination, l'ancien identifiant est envoyé à une table d'archives et le nouvel identifiant remplace l'ancien dans la base de données. Ainsi, un utilisateur qui utilise l'ancien identifiant ou l'ancienne appellation peut facilement retrouver son gène et être informé du nouvel identifiant ou de la nouvelle appellation à utiliser.

Après avoir illustré la démarche de mise à jour, nous avons voulu mesurer l'efficacité de cette dernière.

[5]inconnu en Anglais.

1. Extraction du fichier binaire de Entrez Gene
2. Conversion du fichier binaire en XML
3. Pour chaque gène vérifier si le gène existe dans Ge-Di
   (a) Si le gène existe
      i. Si une mise à jour est requise :
         - Insertion des données manquantes
         - Si l'identifiant n'est plus valide :
            A. Changement de l'identifiant du gène
            B. Archivage de l'ancien identifiant
   (b) Si le gène n'existe pas
      i. Création d'un identifiant
      ii. Insertion des données du gène

TAB. 4.2 – **Le processus de mise à jour du dictionnaire :** Pour chaque gène, l'outil de mise à jour va vérifier si le gène existe dans la base de données. Si ce n'est pas le cas, une insertion est effectuée. Si le gène existe, l'outil doit vérifier si une mise à jour de l'entrée n'est pas requise et que l'identifiant du gène ne doit pas être modifié suite à ces modifications.

4.6 Comparaison

Dans cette section, nous mesurons l'efficacité de notre démarche de mise à jour. Initialement, nous avons voulu que la comparaison englobe la souris. Malheureusement, MGI ne propose à cette date aucun moyen pour télécharger le contenu de sa base de données (Son contenu nous provient à travers Entrez Gene). Par conséquent, nous nous concentrerons sur une seule espèce l'humain. Notre comparaison englobera donc notre plate-forme Ge-Di, Entrez Gene et HGNC.

4.6.1 Tests

Nous avons longuement réfléchi à la manière de comparer clairement des bases de données. Finalement, nous avons trouvé que toute base de données pouvait être représentée comme un ensemble. Ainsi toute comparaison avec un autre ensemble

est simplement l'application de certaines opérations sur les ensembles. Si A et B représentent des ensembles finis, alors nous nous intéresserons tout particulièrement aux ensembles suivants :

- $A \cap B$
- $A \setminus B$ (c'est-à-dire $\{ x \ (x \in A) \wedge (x \notin B) \}$)
- $B \setminus A$ (c'est-à-dire $\{ x \ (x \in B) \wedge (x \notin A) \}$)

Nous utiliserons les diagrammes de Venn pour représenter les ensembles précédents. Ces diagrammes se prêtent tout particulièrement à montrer l'emplacement de l'information dans les ensembles. Nous avons divisé notre comparaison en trois étapes :

Étape 1 : Nous comptabilisons les gènes dans chacune des bases de données. Nous utilisons cette information pour dénombrer les gènes partagés entre les différentes bases de données et ceux propres à une base de donnés en particulier.

Étape 2 : Nous considérons les gènes partagés de l'étape précédente et nous mesurons les conflits concernant le nom et le symbole officiel. Cette étape vise en particulier à mesurer si notre plate-forme est à jour pour le nom et le symbole officiel et si ce n'est pas le cas, nous vérifions si c'est également le cas pour Entrez Gene.

Étape 3 : Nous considérons encore une fois les gènes partagés de l'étape 1 mais cette fois nous extrairons les alias dans chacune des bases de données. Nous dénombrons les alias qui sont partagés ou propres à l'une ou à l'autre des bases de données.

Cette comparaison en trois étapes a pour objectif d'évaluer l'efficacité de notre mise à jour par rapport à la conservation de l'information mais également par rapport aux autres démarches adoptées par les autres bases de données (c'est-à-dire HGNC et Entrez Gene). Nous tenterons par la même occasion de déterminer leurs démarches.

4.6.2 Résultats

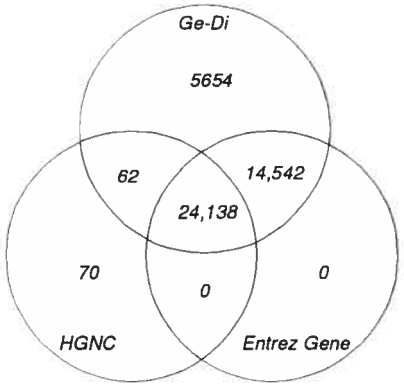

FIG. 4.3 – **Comparaison des gènes contenus dans les différentes bases de données :** En comparant les bases de données, nous avons trouvé que 24,138 gènes sont présents dans les trois bases de données. Entrez Gene et Ge-Di possèdent 14,452 gènes qui ne sont pas contenus dans HGNC. Ces gènes représentent la proportion des gènes connus qui n'ont pas encore été approuvés par HGNC. Finalement, HGNC et Ge-Di se retrouvent à contenir des gènes qui ont été supprimés de Entrez Gene. Ce dernier ne donne aucune explication à ces suppressions.

Dans l'étape 1, nous avons trouvé respectivement 38,680, 44,480 et 24,270 gènes dans Entrez Gene, Ge-Di et HGNC. Le nombre de gènes dans HGNC n'est pas étonnant étant donné son fonctionnement. Sa base de données ne contient que les gènes dont son comité a approuvé le nom et le symbole officiel. Cette différence est clairement visible dans le diagramme de Venn (Figure 4.3). Dans cette figure, nous remarquons que 24,138 gènes sont présents dans les trois bases de données (HGNC $\cap$ Entrez Gene $\cap$ Ge-Di). Ce chiffre est assez proche du nombre de gènes dans HGNC. Ceci nous paraît normal étant donné que les bases de donnés sont connectées indirectement. Entrez Gene télécharge régulièrement le contenu de HGNC et Ge-Di celui de Entrez Gene. Cependant, nous pensions avant les tests que Entrez Gene et Ge-Di allaient contenir strictement tous les gènes de HGNC. Or, nous avons trouvé que HGNC contient 70 gènes propres et 62 qu'il partage avec Ge-Di. Nous avons voulu comprendre la raison de ces écarts et avons donc lancé une étude supplémentaire. Cette étude a consisté à prendre certains de ces gènes et de tenter de retrouver leur trace dans Entrez Gene. Si les 70 gènes propres

de HGNC peuvent ne pas avoir existé dans Entrez Gene, ce n'est pas le cas pour le 62 gènes partagés[6].

Les 70 gènes propres se sont avérés être soit des gènes nouvellement découverts soit des gènes supprimés de Entrez Gene. C'est le cas du gène *C9orf63*[7] qui a été retiré de Entrez Gene depuis le 10 mai 2005. Malheureusement, aucune explication n'est donnée sur sa suppression. Les 62 gènes partagés entre HGNC et Ge-Di sont également des gènes qui ont été supprimés de Entrez Gene sans explication. Il existe une différence cependant, les suppressions ont eu lieu dans les six derniers mois ce qui explique leur présence dans Ge-Di.

Les 5,654 gènes propres à Ge-Di tombent également dans la catégorie des gènes qui ont été supprimés de Entrez Gene. Cependant, ces gènes proviennent d'une source autre que HGNC ce qui explique leur absence de cette dernière. Entrez Gene ne précise pas cette source encore moins la raison des suppressions. Nous aborderons ce problème dans la discussion.

Comme prévu, nous retrouvons qu'il existe une large portion de gènes absents de HGNC. Il s'agit de 14,452 gènes présents dans Entrez Gene et Ge-Di. Ces gènes représentent la proportion des gènes connus qui n'ont pas encore été approuvés par HGNC. Finalement, Entrez Gene n'a pas de gènes propres. Ce dernier résultat ne nous étonne pas. Tout nouveau gène de Entrez Gene s'ajoute à Ge-Di lors de la mise à jour de la plate-forme.

Dans l'étape 2, nous avons extrait les gènes en commun dans les 3 bases de données et nous avons vérifié si le nom et le symbole étaient le même dans chacune des bases de données. Notre référence dans cette étape est HGNC. Ce dernier est chargé d'assigner un nom et un symbole officiel à chaque gène humain. Dans Ge-Di, nous avons trouvé 50 gènes (0.2 % des 24,138 gènes) qui ne possèdent pas le nom et le symbole officiel. Entrez Gene quant à lui, ne possède que 43 gènes (0.17 % des 24,138 gènes) dans cette situation. La différence entre Entrez Gene et Ge-Di est minime au vu des pourcentages.

---

[6]Tout gène qui est contenu dans Ge-Di provient à l'origine de Entrez Gene

[7]HGNC ID : 26093

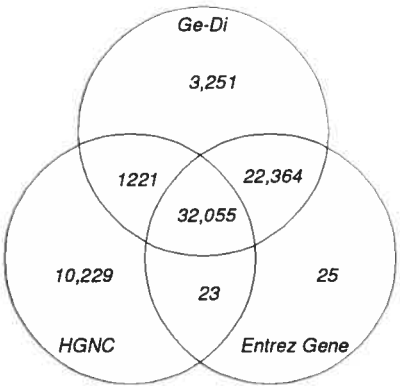

FIG. 4.4 – **Comparaison des alias contenus dans les différentes bases de données :** Dans notre comparaison, nous avons trouvé 32,055 alias en commun pour les 24,138 gènes en commun. Pour les alias présents dans les deux bases de données, Entrez Gene et Ge-Di prennent les premières places. Ce résultat est peut être une preuve que HGNC n'est pas le seul fournisseur de Entrez Gene. Le couple Ge-Di/HGNC suit de près avec 1221 alias. Ce chiffre représente probablement des alias qui ont été perdus par Entrez Gene. Finalement, HGNC et Entrez Gene ferment la marche avec 23 gènes. Pour les alias contenus seulement dans une base de données, Ge-Di et HGNC prennent les premières places. Les alias propres à HGNC représentent soit des nouvelles appellations soit des anciennes appellations. Nous ignorons cependant la raison de l'absence de ces anciennes appellations dans Entrez Gene. Les alias propres à Ge-Di quant à eux, représentent des appellations qui ont été perdues par Entrez Gene au cours du temps.

Dans l'étape 3, nous avons également considéré les 24,138 gènes et extrait les alias pour chacun de ces gènes. Par la suite, nous avons simplement déterminé les alias en commun et les alias propres. Nous avons trouvé 32,055 alias communs (Figure 4.4). Pour les alias présents strictement dans deux bases de données, le couple Ge-Di/Entrez Gene est en tête avec 22,364. Nous n'avons pas pu trouver d'explications à l'absence de ces alias dans HGNC. Il semblerait que Entrez Gene ait plus qu'un fournisseur pour l'information des gènes chez l'humain. Mais ceci n'est qu'une hypothèse dont nous parlerons dans la discussion. Le couple HGNC/Ge-Di arrive en deuxième position avec 1221 alias. Il semblerait que Entrez Gene ait perdu ces alias sinon ils n'existeraient pas dans Ge-Di. Le couple HGNC/Entrez Gene arrive en dernière position avec 23 alias. Ces derniers représentent des alias

qui ont été récemment rajoutés aux deux bases de données.
Pour les alias propres, HGNC dépasse les autres bases de données avec 10,229 alias. Nous avons voulu comprendre l'importance de ce chiffre. Nous avons encore une fois lancé un étude complémentaire et nous avons trouvé que certains de ces alias sont de nouveaux alias qui ont été rajoutés récemment et d'autres dont l'ajout n'est pas récent. Nous ignorons cependant les raisons de leur absence de Entrez Gene.
Finalement, Ge-Di arrive en deuxième place avec 3251 alias propres. Ces derniers représentent des noms et symboles de gènes qui ont été perdus de Entrez Gene. Ce sont pour la plupart des appellations non officielles qui ont été écrasées par les noms et symboles officiels. Entrez Gene ferme la marche avec 25 alias propres. Ces derniers représentent des nouvelles appellations de gènes qui ont été rajoutées récemment et qui vont être rajoutées à la plate-forme lors de la prochaine mise à jour.
Finalement, lors de cette étude, nous avons trouvé également des alias qui existent dans les trois bases de données dont la capitalisation diffère c'est le cas de l'alias *Kaiso*[8]. Ce dernier apparaît en tant que *Kaiso* dans Entrez Gene et *KAISO* dans HGNC. Nous avons songé à recommencer cette étape à cause de cette différence dans l'orthographe mais nous avons choisi de ne pas le faire. La différence n'est pas négligeable car elle peut faire échouer un outil d'identification de gène dans un texte donné. Cet argument nous a convaincu de la validité de nos résultats.

4.6.3 Discussion

La biologie est probablement une des sciences les plus dynamiques. La comparaison que nous avons effectuée atteste de la rapidité et de la quantité des changements. En moins de 6 mois, 5,654 gènes ont été supprimés de Entrez Gene. Ces changements n'affectent pas seulement l'ensemble de l'information des gènes mais également leurs appellations. Cependant, nous ne possédons pas toutes les informations possibles pour quantifier les changements sur l'information d'un gène mais

[8]Entrez Gene ID :10009. HGNC ID :16682

nous pensons qu'ils sont assez importants au vu des 3,251 alias qui ont disparu de Entrez Gene et qui ont été sauvegardés par notre procédure de mise à jour. Ces changements soulignent l'importance de la mise à jour en particulier dans des domaines aussi dynamiques que la biologie.

Dans la première étape de notre test, nous avons remarqué des différences entre le nombre de gènes présents dans les bases de données. HGNC se trouve désavantagé cependant ce n'est qu'un inconvénient temporaire. Le but de HGNC est de couvrir tout le génome humain et nous pensons que ce comité arrivera à le faire. Revenons aux chiffres et concentrons nous sur notre plate-forme. Dans la comparaison, Ge-Di se retrouve avec 5654 gènes propres. Comme nous l'avons mentionné auparavant, il s'agit des gènes qui apparaissent avec la mention *"This record was discontinued."* dans Entrez Gene. Nous nous ne pouvons pas vraiment considérer ce chiffre comme un avantage mais plutôt comme un inconvénient dans la procédure de mise à jour. Cette dernière comme les comparaisons de la section précédente exclue les gènes obsolètes (c'est-à-dire ceux avec la mention discontinued). Sans cette exclusion, nous aurions 100,000 gènes à traiter au lieu des 40,000 approximativement. Nous avons choisi au début d'exclure les gènes obsolètes pour éviter la redondance et de partir sur une bonne base. Plus tard, nous avons voulu développer des outils pour prendre en charge ces gènes devenus tout d'un coup obsolètes mais malheureusement l'information manque. Pour pouvoir développer un outil, il faut être capable d'effectuer la tâche manuellement. Or, Entrez Gene ne précise pas toujours la raison de la suppression. Nous avons décidé de ne pas exclure ces gènes obsolètes et d'attendre que Entrez Gene donne plus d'explications dans ses changements. En attendant, nous considérons que toute information est bonne qu'elle soit obsolète ou pas. De plus, nous ignorons le nombre d'articles qui mentionnent ces gènes obsolètes. Nous ne voulons pas prendre le risque d'être un frein à la recherche.

La deuxième partie du test marque encore une fois le dynamisme de la biologie ou dans ce cas le dynamique de la formalisation. Dans cette étape, nous remarquons qu'il existe un certain décalage entre la plate-forme et son fournisseur mais également qu'il existe un décalage plus important entre Entrez Gene et HGNC. Ce

décalage souligne peut être la nécessité d'une mise à jour automatique.

La troisième partie est probablement la partie la plus importante de la comparaison. Si dans les deux étapes précédentes, avoir de l'information propre pouvait représenter un problème, ce n'est pas le cas dans cette partie. Dans cette partie, Ge-Di arrive à sauvegarder l'information. Cette remarque est illustrée par ses 3,251 alias propres mais également les 1,221 alias qu'il partage avec HGNC. Ces alias proviennent de Entrez Gene car notre plate-forme utilise uniquement le contenu de Entrez Gene pour les mises à jour. Ces alias représentent pour la plupart des anciens noms et symboles officiels qui ont été remplacés par d'autres appellations plus appropriées. HGNC a probablement compris que le gène quelque soit son appellation représente une information importante et qu'il est nécessaire que l'utilisateur puisse retrouver sa trace dans la littérature. Malheureusement, HGNC ne s'attaque qu'à une seule partie du problème et nous pensons que Ge-Di puisse aider à traiter l'autre partie du problème. Les résultats vont dans ce sens.

Dans la présentation des résultats, nous avons mentionné l'hypothèse que HGNC ne soit pas le seul fournisseur tout au moins pour les alias. Cette hypothèse se base surtout sur les 22,364 alias qui sont partagés entre Entrez Gene et Ge-Di. Si HGNC était le seul fournisseur de Entrez Gene, logiquement une plus grande portion existerait chez HGNC. Or ce n'est pas le cas. Entrez Gene domine avec 58,695 alias. Nous avons songé un certain moment que HGNC aurait pu perdre des alias tout comme Entrez Gene. Cette hypothèse est mise à mal par les 10,229 alias propres. Cette hypothèse est également mise à mal par l'organisation de HGNC. Pour les appellations de gènes, il existe 5 catégories :

- Symbole Officiel
- Nom Officiel
- Anciens Symboles Officiels
- Anciens Noms Officiels
- Alias

Nous avons voulu découvrir cette autre source. Malheureusement, notre démarche a été sans succès. Il faut simplement considérer que pour l'humain, il existe proba-

blement d'autres sources utilisées par Entrez Gene.

De cette comparaison, nous avons souligné certaines qualités de la plate-forme mais nous avons découvert certains défauts. Nous avons déjà discuté d'un des défauts mais nous pouvons également discuté des alias propres à HGNC. HGNC possède 10,229 alias propres. Il est vrai que certains de ses alias viendront se rajouter à Entrez Gene et par la suite à Ge-Di. Cependant, une grande partie néanmoins restera dans HGNC, il serait souhaitable de rajouter cette liste d'alias propres à Ge-Di.

CHAPITRE 5

APPLICATIONS

Dans les chapitres précédents, nous avons décrit les différentes composantes de la plate-forme. Dans ce chapitre, nous tentons de comparer l'ensemble de la plate-forme à des outils existants.

5.1 Une comparaison adéquate

Nous avons longuement réfléchi à la manière d'effectuer la comparaison avec les outils existants et quels critères prendre en compte. Nous n'avons pas trouvé un modèle de comparaison adapté à une plate-forme en général. Les modèles existent cependant ils ne couvrent pas tous les critères que nous souhaitons évaluer. Une plate-forme n'est pas un algorithme dont on peut calculer une complexité et par la suite évaluer la performance. Une plate-forme est plutôt une expérience. Il faut donc évaluer la facilité de la prise en charge et les performances lors d'une utilisation donnée. Nous avons choisi de ne pas évaluer le premier critère car nous pensons que la prise en charge est très subjective. Tout utilisateur peut arriver à maîtriser un outil si ce dernier est bien construit. Nous nous concentrerons uniquement sur le deuxième critère. Nous définissons la performance comme étant le nombre de faux positifs par requête. Plus ce nombre est élevé, moins l'outil est performant. Nous sommes encore dans le contexte des réponses courtes c'est-à-dire répondre à la question de l'appartenance d'un gène à un organisme donné ou trouver l'information du gène.

Nous avons voulu au départ automatiser la comparaison pour avoir un ensemble assez significatif statistiquement. Malheureusement, cette automatisation n'a pas été possible. La plupart des outils concurrents ne permettent plus ou ne permettent pas des recherches à partir d'applications locales. C'est le cas de Google. Ce dernier

proposait une interface de programmation[1] pour les applications locales et en ligne. Cette interface n'est plus disponible depuis le 5 décembre 2006. Elle a été remplacée par une interface de programmation orientée uniquement vers les applications en ligne.

Notre comparaison est divisée en deux parties. Dans la première partie, nous évaluons la performance de l'interface graphique et nous la comparons avec : Google, Entrez Gene, HGNC et *Gene Cards* [RCCPL97]. Dans la deuxième partie, nous évaluons l'interface de programmation. Nous comparons cette dernière avec l'interface de programmation de Entrez Gene.

5.2 Interface Graphique

Nous avons divisé la comparaison de l'interface graphique en plusieurs petits tests. Chaque test représente la mesure d'un critère donné :

- Recherche de gènes connus
- Recherche de gènes ambigus (homonymes)
- Recherche approximative de gènes

5.2.1 Recherche de gènes connus

Dans ce test, nous tentons de retrouver l'information pour 43 gènes. Ces gènes ont été extraits de Wikipedia[2]. Ils représentent les gènes les plus connus actuellement. Cette popularité peut être expliquée par leur implication dans certaines maladies mais également dans certaines controverses (Le gène *ZBTB7A* se nommait *pokemon*). Cette notoriété nous permet de penser sans crainte que ces gènes sont mentionnés dans un grand nombre d'articles. Par conséquent, aucun outil n'est défavorisé.

Nous avons défini certaines procédures pour ce test ainsi que pour ceux qui suivent. La première procédure est la limitation du nombre de résultats à traiter par entrée.

[1] Google SOAP Search API : http ://code.google.com/apis/soapsearch/
[2] http ://en.wikipedia.org/wiki/List_of_notable_genes

i	Symbole	Google	*Entrez Gene*	*HGNC*	*Ge-Di*	*Gene Cards*
1	ALB	11	1	3	0	0
2	BCL2	1	4	11	0	0
3	CCR5	2	1	1	0	0
4	CD4	6	1	2	0	0
5	CD8	-	1	-	0	0
6	IL2	3	1	6	0	0
7	IL10	1	1	2	0	0
8	BRCA1	2	1	4	0	0
9	BRCA2	2	1	3	0	0
10	CD28	3	2	1	0	0
11	ZBTB7A	1	1	1	0	0
12	APC	82	2	9	0	0
13	ASPM	3	1	2	0	0
14	BDNF	3	3	1	0	0
15	CFTR	3	1	14	0	0
16	CREBBP	1	1	1	0	0
17	CRH	12	2	2	0	0
18	CXCR4	1	1	1	0	0
19	DHFR	4	2	1	0	1
20	HFE	-	1	3	0	1
21	KRT14	2	1	3	0	0
22	KRT5	2	1	2	0	0
23	PGL2	8	1	2	0	0
24	RHO	38	-	98	1	1
25	SDHB	2	-	2	0	0
26	SDHC	11	11	2	0	0
27	SDHD	1	9	2	0	0
28	SRY	-	13	22	0	1
29	TSC1	1	3	1	0	1
30	TSC2	1	2	2	0	0
31	APP	28	1	-	0	0
32	GAST	15	2	4	0	0
33	LCK	2	1	2	0	0
34	LEP	37	4	31	0	0
35	LIF	-	2	23	1	0
36	MCM6	5	1	1	0	0
37	MYH7	1	1	3	0	0
38	MYOD1	6	2	2	0	0
39	NPPB	0	4	1	0	0
40	OSM	21	3	26	0	0
41	PKC	-	-	-	0	0
42	PIP	88	16	17	2	1
43	SLC18A2	1	3	2	0	0
	Moyenne	10.82	2.75	7.90	0.09	0.14
	Écart Type	20.09	3.44	16.50	0.37	0.35
	Médiane	3	1	2	0	0

TAB. 5.1 – **Recherche de gènes connus :** Nous avons extrait de Wikipedia la liste des gènes les plus populaires actuellement. Nous avons tenté de retrouver l'information pour chacun de ces gènes. Les chiffres mentionnés représentent le nombre de faux positifs qui précédent l'information pertinente. Gene Card Et Ge-Di se sont avérés être les plus performants avec des moyennes respectives de 0.14 et 0.09 faux positifs par requête. Un tiret signifie que l'information du gène n'a pas été trouvée dans les 100 premiers résultats.

Nous considérons qu'un utilisateur ne va pas consulter plus de 100 résultats par entrée. S'il ne trouve pas son information dans les 100 premiers résultats, il va reformuler son entrée. Dans ce test, reformuler l'entrée signifie que l'outil a échoué et nous représentons cet échec par un tiret. La deuxième procédure est reliée au nombre généré de résultats. Nous dénombrons uniquement le nombre de faux positifs qui précède l'information souhaitée. Il se peut très bien qu'un outil génère un nombre impressionnant de faux positifs mais l'information souhaitée apparaît en premier. Le cas idéal est de se retrouver avec l'information directement (c'est-à-dire 0 faux positif avant l'information souhaitée).

Le résultat de ce test est présenté dans le tableau 5.1. Durant ce test, le moteur de recherche Google nous a fait découvrir un nouvel outil. Il s'agit de *Gene Cards*. Les performances de cette base de données nous ont particulièrement impressionné et nous avons décidé de l'inclure dans ce test et dans les suivants. À première vue, les résultats semblent prouver qu'il est plus utile d'utiliser un moteur de recherche spécialisé plutôt que Google. Néamoins, nous estimons que la puissance de Google n'est pas remise en cause. Il ne permet pas de retrouver l'information recherchée pour 5 gènes (11.63 % des gènes recherchés) mais à titre de comparaison HGNC et Entrez Gene ne font pas mieux avec 3 gènes (6.98 % des gènes recherchés). C'est plutôt à cause des faux positifs que le moteur de recherche le plus connu se retrouve en dernier. Il génère une moyenne de 10 faux positifs par requête avec un écart type de 20. Pratiquement, cette moyenne et cet écart type se traduisent par de grandes disparités pour les requêtes. L'utilisateur va se retrouver avec soit beaucoup de résultats soit avec très peu. Ce résultat n'est guère étonnant, le but de Google est de couvrir toutes les pages Web et non pas uniquement les pages mentionnant l'information des gènes. Google a cependant confirmé à nos yeux son rang et son utilité. Il nous a permis de découvrir un outil dont nous ignorons l'existence : Gene Cards.

Les résultats de HGNC nous ont d'une certaine manière étonné. HGNC avait un grand avantage sur les autres outils. Sa base de données ne contient que l'information des gènes pour l'humain. Or la liste de gènes que nous recherchons est

constituée strictement de gènes humains. Malgré cet avantage, HGNC a une performance similaire à Google. Entrez Gene quant à lui, nous a donné des résultats plus que satisfaisants. Cependant, nous avons été déçus de ne pas retrouver de l'information pour 3 gènes. L'information existe. Cependant, le moteur de recherche de Entrez Gene n'est pas sensible à la casse. Ce défaut malheureusement, explique cet échec.

Finalement, le couple Ge-Di/Gene Cards nous a particulièrement impressionné. Le deux outils ont permis de retrouver l'information des gènes avec une moyenne de 0.09 et 0.14 faux positifs c'est à dire pratiquement aucun faux positif. Ce résultat est important car il prouve que ces deux outils permettent un gain de temps considérable par rapport aux autres. Dans le monde de la recherche, le temps est un facteur crucial. Il nous semble évident qu'un utilisateur va toujours choisir un outil équivalent qui prend moins de temps. Dans cette perspective, Ge-Di et Gene Card semblent être en bonne position pour accélérer la recherche de l'information des gènes en général.

Dans ce test, nous avons pris les différents outils avec leurs réglages par défaut. Nous considérons que l'utilisateur va choisir les réglages par défaut avant de tenter de les modifier. Pour Google, nous avons choisi de ne pas spécifier de mots-clefs (tel que *gene*[3]). En théorie, ce choix peut le défavoriser. En pratique, ce choix n'a pas été fatal. Mais si nous l'effectuons, nous devrions également changer les réglages de chacun des outils et surtout juger quelles modifications sont équivalentes. Garder une certaine égalité entre les différents outils nous a paru très difficile dans ce cas. Par conséquent, nous avons renoncé à ce choix dans ce test et dans les suivants. Cependant, nous mentionnons si une modification de l'entrée (comme l'ajout de mots-clefs) ou des variables du moteur améliorent les résultats.

[3]Gène en anglais

5.2.2 Recherche d'homonymes

Nous avons souvent discuté dans ce mémoire des appellations de gènes ambiguës ou homonymes. En général, ces derniers retardent l'accès à l'information. Si un homonyme apparaît dans la littérature, l'utilisateur doit vérifier de quel gène il s'agit avant de tenter d'extraire son information dans une base de données. Une fois cette étape terminée, il peut tenter d'extraire son information dans une base de données. Mais encore une fois, cette étape doit être suivie d'une étape de vérification. Pour chacun des résultats, l'utilisateur doit vérifier si cette information correspond bien au gène en question. En général, nous pouvons classer les appellations ambiguës dans deux catégories distinctes[4] :

- Les appellations de gènes correspondants à plus d'un gène dans une espèce
- Les appellations de gènes correspondants à plus d'un gène dans plus d'une espèce

Dans ce test, nous visons à mesurer si les différents outils permettent d'aider l'utilisateur à réduire l'ambiguïté du gène recherché. Nous savons que le problème ne peut pas être résolu mais en pratique certaines mesures peuvent être mises en place. Ces mesures visent à réduire l'ambiguïté ou simplement informer l'utilisateur de l'ambiguïté du terme recherché. À titre d'exemple, l'encyclopédie en ligne Wikipedia[5] propose déjà un outil qui demande à l'utilisateur des précisions sur sa requête si cette dernière est ambiguë.

Nous avons par conséquent extrait de notre base de données 48 appellations qui correspondent à plus d'un seul gène dans l'humain[6] et nous avons simplement mesuré si les utilisateurs étaient au moins informés de l'ambiguité de ces noms. Encore une fois, nous plaçons une limite dans le nombre de résultats par entrée. Cette limite est de 100 résultats.

---

[4] À noter que certains noms de gènes peuvent exister dans les deux catégories (par exemple *HOX4*).

[5] http ://www.wikipedia.org

[6] Nous avons choisi de ne pas exclure Gene Card de ce test. Gene Card ne contient que les gènes humains.

i	nom de gène	Google	Entrez Gene	Gene Card	Ge-Di
1	A1B	0	2	2	2
2	ACT	0	7^+	7	7
3	DAC	0	2^+	2	2
4	GP110	0	2	2	2
5	ABC1	1	1^+	2	2
6	B29	1	2	2	2
7	RMP	0	3	3	3
8	ABL	0	2^+	2	2
9	p150	0	5^+	6	4
10	ABP	0	1^+	2	2
11	DAO	1	1	2	2
12	ARG	0	2^+	2	2
13	GTB	0	2	2	2
14	A3GALNT	2^+	2	2	2
15	ACCA	0	0	2	2
16	P58	0	7^+	7	3
17	MCAD	0	2	2	2
18	SCAD	0	1	1	2
19	T2	0	1	2	2
20	MAT	0	2^+	2	2
21	ACAT	0	3	3	3
22	BNC1	0	2	2	2
23	CLK-1	0	2	2	2
24	CDC25	0	2	2	2
25	TRAP	0	3^+	5	4
26	PAP	0	5^+	8	8
27	NEM1	0	2	2	2
28	NEM2	1	2	2	2
29	BWS	0	2	2	2
30	WBS	0	3	3	5
31	CDH14	1	2	2	2
32	CAK1	0	2	2	2
33	ACTL6	2	2	2	2
34	FOP	0	2	3	3
35	MLM	0	2	2	2
36	P57	0	2	2	2
37	CIP1	0	2	2	2
38	ACVRLK4	2	2	2	2
39	ALK1	1	2	2	2
40	AN	0	0	4	3
41	p14	0	2	9	3
42	STK1	1	2	2	2
43	OB	0	2^+	2	2
44	MADM	1	2^+	2	2
45	DSH	0	2^+	2	2
46	G1P1	1	2	1	2
47	P21	0	2^+	3	2
48	KIP2	0	2	2	2

TAB. 5.2 – **Recherche de gènes ambigus :** Google et Entrez Gene ne permettent pas de retrouver les différents gènes qui correspondent à un homonyme. De plus, ces deux outils n'informent pas l'utilisateur de l'ambiguïté du terme recherché. Ce n'est pas le cas pour Ge-Di et Gene Card. Gene Card se montre également très performant sur le nombre de gènes retrouvés pour un terme ambigu donné. Dans ce tableau, nous dénombrons le nombre de gènes trouvés dans les 100 premiers résultats. Le symbole '+' dénote les entrées qui ont nécessité le traitement de plus de 10 résultats.

Dans ce test, nous ne dénombrons pas le nombre de faux positifs mais plutôt le nombre de gènes qui peut être trouvés dans les 100 premiers résultats. Nous faisons également une différence entre les gènes trouvés dans les 10 premiers résultats et ceux qui sont trouvés dans 90 résultats restants. Cette différence est importante car elle permet un gain de temps considérable. Les résultats de ce test sont présentés dans le tableau 5.2. Ce test marque une grande différence entre les outils qui informent les utilisateurs de l'ambiguïté et ceux qui ne le font pas. Google et Entrez gene se trouvent dans la dernière catégorie. Nous avons déjà mentionné dans ce mémoire que ces deux outils généraient un nombre assez important de résultats. Dans le domaine de l'ambiguïté, ces deux outils ne permettent pas à l'utilisateur de découvrir l'ambiguïté du terme recherché. Le seul moyen de découvrir l'ambiguïté d'un terme donné est de traiter plus de résultats pour chacune des requêtes. Mais pourquoi devrait-on le faire si nous possédons déjà certains résultats pertinents ? Nous ne sommes pas non plus en mesure de déterminer le chiffre exact de résultats à traiter et nous ne pensons pas qu'un tel chiffre puisse exister. Nous savons néanmoins que les utilisateurs ne vont pas traiter les milliers ou les centaines de résultats [Mor05]. En pratique, les utilisateurs consultent que les premiers résultats. Ces derniers sont les plus pertinents selon eux. C'est ainsi que la plupart des gens utilisent Google [Mor05]. Lorsqu'il ne donne pas de résultats pertinents, rajouter des mots clefs semble être une solution consensus dans le domaine de l'informatique. Nous avons tenté de rajouter des mots clefs pour améliorer ses résultats. Cette solution n'a pas vraiment eu les résultats attendus. Elle a permis parfois de retrouver certains résultats lorsqu'aucun n'était généré sans mots-clefs. Cependant, l'information retrouvée est en général une publication ou un article mais rien qui puisse informer l'utilisateur d'une ambiguïté potentielle.

L'ordre d'apparition des résultats souligne également la différence entre les deux catégories. 95.83 % et 31.25% des gènes recherchés respectivement dans Google et Entrez ont requis le traitement de 10 résultats ou plus pour chacun des gènes. Pour Ge-Di et Gene Card, aucun des gènes a requis un tel traitement, les deux outils informent directement l'utilisateur de l'ambiguïté. Ce moyen simple est de

plus en plus proposé dans la littérature comme un moyen de réduire l'ambiguïté [WSVM+03]. L'autre moyen complémentaire est de forcer les auteurs dans la littérature à clarifier le gène mentionné en rajoutant des informations supplémentaires (comme un identifiant de base de données). C'est peut être le constat que le problème ne peut pas être totalement résolu. Néanmoins, nous pensons que certains moyens peuvent être mis en place en pratique pour aider l'utilisateur. Ces dernières ne permettent pas uniquement un gain temporel. Informer l'utilisateur de l'ambiguïté d'un terme peut l'aider à éliminer certains faux positifs d'un ensemble à traiter.

Finalement, les performances des différents outils sont presque similaires à ceux du test précédent. Google se retrouve encore une fois dernier. Il arrive au même niveau des autres outils que 3 fois (6.25 %). Il est suivi de loin par le trio de tête. À la troisième place, nous retrouvons Entrez Gene. Ce dernier a permis de retrouver 35 fois l'ensemble des gènes pour un homonyme donné (72.92 %). La deuxième place revient Ge-Di. Notre plate-forme arrive à trouver les gènes 41 fois (85.47 %). Finalement la première place revient à Gene Card (44 fois - 91.67 %). Le but de ce test n'était pas de mesurer le contenu de chacune des bases de données mais plutôt les moyens mis en place pour réduire l'ambiguïté. Sur ce critère, Ge-Di et Gene Card excellent. Entrez Gene reste satisfaisant mais il est un peu en recul.

Nous avons lancé une analyse supplémentaire. Nous avons voulu savoir si un outil s'était démarqué pour le contenu. Cette place revient encore une fois à Gene Card. Ce dernier permet d'identifier plus de gènes par homonyme lors de 6 requêtes (12.50 %). Il est encore une fois suivi par Ge-Di avec 3 requêtes (6.25 %). Google et Entrez Gene n'ont produit aucun résultat de ce type. Pour la comparaison Entrez/Ge-Di, Ge-Di génère plus de gènes par requête 4 fois. Entrez Gene se retrouve dans cette position que deux fois. Nous pensons que les gènes manquants dans la plate-forme seront probablement rajoutés à la prochaine mise à jour. Quant à ceux de Entrez Gene, nous pensons qu'ils tombent dans la catégorie des informations perdues (voir section 4.6.2).

Avant de clore ce test, nous avons voulu lancer une dernière comparaison. Nous avons discuté dans la section 1.5 que MGI et HGNC tentaient d'assigner le même

symbole aux gènes équivalents chez l'humain et la souris. Cependant, les lettres majuscules permettent de différencier le symbole humain et celui de la souris. Malgré cette différence, certains auteurs parlent d'ambiguïté entre les deux espèces. Nous avons voulu voir si les démarches de HGNC et MGI ont compliqué la recherche des gènes avec Google ou Entrez Gene. Nous avons donc extrait de notre de base de données 25 gènes équivalents (Tableau 5.3) et tenté de les retrouver en utilisant ces deux outils. Les résultats ont confirmé nos craintes. Nous avons trouvé que même si l'utilisateur respecte les démarches de formalisation, il risque de tomber sur le gène d'une autre espèce. Ce problème est particulièrement courant chez la souris. En utilisant Google, 86.36 % des requêtes pointent en premier vers les gènes humains plutôt que ceux de la souris. Chez l'humain, seulement 4.54 % des requêtes sur un gène humain ont abouti sur un gène de la souris. En utilisant Entrez Gene, le résultat est un peu différent. Les proportions sont respectivement 52 % et 36 % chez la souris et l'humain. Ce résultat prouve qu'il existe également une confusion pour les symboles communs de la souris et l'humain. Cependant, cette confusion n'est pas vraiment à cause d'une mauvaise organisation des biologistes. Les concepteurs des moteurs de recherche ont choisi de ne pas proposer des moyens de régler la sensibilité à la casse. Nous pensons sérieusement que ce choix est un grand désavantage pour les utilisateurs. De plus, ces derniers ne connaissent pas toujours les gènes recherchés. Le fonctionnement de la biologie actuelle est tel que le biologiste doit souvent aller rechercher l'information en dehors de son domaine d'expertise [JSB06]. Les outils de recherche dans ce cas de figure, font en sorte de rendre difficile l'apprentissage de nouvelles connaissances. À chaque résultat généré, l'utilisateur doit vérifier s'il s'agit bien d'un gène de la souris ou d'un humain. Malheureusement, notre plate-forme permet uniquement de retrouver l'information du gène et non pas la littérature correspondante.

5.2.3 Recherche approximative de gènes

Nous avons discuté dans la section 3.1 du nombre impressionnant de gènes connus. Ce nombre implique que les appellations de gènes ne sont pas toujours

i	Symbole Humain	Symbole de la souris
1	ACP1	Acp1
2	AOAH	Aoah
3	ARSB	Arsb
4	SERPINC1	Serpinc1
5	BMP	Bmp
6	C1QC	C1qc
7	CACNAF1	Cacnaf1
8	CAPG	Capg
9	RUNX2	Runx2
10	CD247	Cd247
11	ENTPD2	Entpd2
12	CEL	Cel
13	DCN	Dcn
14	ELK3	Elk3
15	GAA	Gaa
16	GEM	Gem
17	JARID2	Jarid2
18	MAGEA9	Magea9
19	NONO	Nono
20	POU4F2	Pou4f2
21	PTN	Ptn
22	RET	Ret
23	CORO2A	Coro2a
24	SLN	Sln
25	AURKA	Aurka

TAB. 5.3 – **Liste de gènes équivalents dans la souris et l'humain :** Les comités de HGNC et MGI ont choisi d'assigner aux gènes équivalents dans les deux espèces le même symbole tout en respectant les majuscules et les minuscules (voir section 1.5). Cette démarche de formalisation rend difficile la recherche d'une espèce donnée sur Internet à cause du manque de sensibilité des moteurs de recherche.

connues avec précision. Dans ce test, nous visons à évaluer si une approximation de l'entrée est fatale à la recherche d'information. Ici, le terme approximation est pris dans un sens particulier. Dans cette section, nous considérons les entrées incorrectement orthographiées. Ces dernières ressemblent *approximativement* à celles correctement orthographiées. Nous avons réfléchi au moyen de générer ces approximations d'une façon automatique. Une page de wikipedia nous a grandement inspiré[7]. Cette dernière ne fait que lister les erreurs les plus connues de tape. Nous avons remarqué dans cette liste deux types d'erreurs. Le premier type est le remplacement d'un caractère par un autre (manger → mangdr). Le deuxième type est le changement de l'ordre d'apparition des caractères dans une chaîne donnée (manger → amnger). Nous avons par conséquent décidé de construire un programme

[7]http ://en.wikipedia.org/wiki/Lists_of_common_misspellings

qui génère ces deux types d'erreurs de manière aléatoire. Ces erreurs ne donnent pas vraiment un grand avantage à notre outil par rapport aux autres. Par besoin de clarté, nous avons nommé le premier type d'erreurs *substitution* et le deuxième type d'erreurs *réarrangement.*

Pour le test de substitution, les différents outils nous ont particulièrement déçu (Tableau 5.4). Sur les 30 appellations incorrectement orthographiées, seulement 2 gènes ont pu être retrouvés (6.67 %). C'est Entrez Gene qui permet de retrouver ces deux gènes. Pour une fois, Google n'est pas dernier et permet de retrouver un seul gène (3.33 %). Gene Card se retrouve donc à la dernière place avec aucun gène retrouvé. La première place revient à Ge-Di. Notre plate-forme arrive à retrouver 18 gènes (60 %). Ce résultat est un peu étonnant car les erreurs générées ne favorisent en aucun cas notre algorithme de suggestion. Cependant, notre outil génère parfois un nombre excessif de résultats. Nous discuterons plus tard de ce problème. Revenons aux autres outils et essayons de comprendre leur échec. Pour Google, la suggestion se fait en comparant le nombre de résultats. Si un terme similaire dans l'orthographe génère plus de résultats alors Google va faire une suggestion. Dans ce test, nous pensons que les appellations de gènes correctement orthographiées ne génèrent pas plus de résultats. Par conséquent, Google n'effectue aucune suggestion. Entrez Gene possède également un outil de suggestion. Il s'est avéré plus performant que Google mais bien loin de celui de Ge-Di. Finalement, Gene Card doit son échec à l'absence d'un outil de suggestion dans son moteur de recherche.

Pour les réarrangements, la performance des différents outils est en baisse (Tableau 5.5). Si le test précédent a permis aux différents outils de retrouver certains gènes, Ici, seulement notre plate-forme permet de retrouver certains gènes. Sa performance est de l'ordre du 50 %. Ce dernier test a permis de prouver notre remarque sur les moteurs de recherche en général. Ces derniers supposent presque toujours que l'entrée est exacte même si aucun résultat n'est généré.

Or en pratique, l'entrée peut être incorrectement orthographiée. Cette remarque est particulièrement véridique pour les gènes. Cependant, il faut noter une certaine nuance sur les approximations. Certaines approximations peuvent représenter des

i	alias/modifié	*Google*	*Entrez Gene*	*Gene Cards*	*Ge-Di*
1	KIAA1232/KIAK1232	-	-	-	12*
2	HTPCRH06/HTBCRH06	-	-	-	1*
3	CLEC15B/YLEC15B	-	-	-	1*
4	MGC9454/MGC2454	100	1	-	1
5	D21S58/D61S58	-	-	-	100
6	FLJ35924/FLY35924	-	-	-	1*
7	OR3.2/OJ3.2	100	-	-	100
8	HSU84971/HSU84991	-	-	-	0*
9	FLJ10307/FLJ130307	24	2	1	1
10	AE2/AE7	100	16	1	9
11	CGN5/CGN2	100	97	-	2
12	MIP/MSP	100	100	-	23
13	KNSL6/SNSL6	8	-	-	6*
14	filamin B/filsmin B	0*	0*	-	26*
15	SMARD1/SMARB1	9	-	-	3
16	NRC/XRC	100	-	-	16
17	SFMBT/SFMBL	100	0*	-	8*
18	WASP/WAKP	100	-	-	100*
19	NAT8BP/NAS8BP	-	-	-	7*
20	DLTA/DUTA	100	-	-	3
21	HOMG2/HOMG3	100	-	-	2*
22	FLJ22251/QLJ22251	100	-	-	1*
23	CTRP5/CTHP5	53	-	-	5
24	B56B/B06B	100	-	-	11
25	IL-21/IY-21	100	-	-	30*
26	OR4-114/WR4-114	2	-	-	10*
27	GLNS/GLNZ	100	-	-	100*
28	DDXL/DDXT	100	-	-	44*
29	HBII-135/XBII-135	100	-	-	12*
30	U15A/L15A	100	1	-	14*
	pourcentage de réussite	3.3 %	6.6 %	0 %	60 %

TAB. 5.4 – **Recherche approximative - substitution :** La recherche des symboles incorrectement orthographiés prouvent que la plupart des moteurs de recherche ne permettent pas de retrouver les gènes en question. Les chiffres représentent les faux positifs. Le symbole '*' représente la réussite de la recherche et '-' signifie qu'aucun résultat n'a été généré.

chaînes valides de caractères (c'est-à-dire des symboles valides dans le même domaine ou dans un autre domaine). 73 % des noms de gènes incorrectement orthographiés ont généré des résultats avec Google. Il ne s'agissait pas nécessairement de gènes mais Google a généré des résultats correspondants à l'entrée. Dans ce contexte, les résultats de Google ne sont pas nécessairement mauvais. Nous découvrons peut être qu'il existe deux nouveaux types de confusions. Le premier type de confusion est entre les symboles incorrectement orthographiés et ceux correctement orthographiés. Le deuxième type de confusion est entre les symboles incorrectement orthographiés et d'autres termes dans d'autres domaines. À cette date, nous pen-

i	alias/modifié	*Google*	*Entrez Gene*	*Ge-Di*	*Gene Card*
1	MGC3178 / MGC**1**378	2	-	24	-
2	p31 / p**1**3	100	10	4	1
3	OEATC-1 / **E**OATC-1	1	-	1*	-
4	FLJ31384 / **L**FJ31384	-	-	1*	-
5	Tiff66 / Tif**6**f6	-	-	9*	-
6	NOS2 / NO**2**S	100	-	100*	-
7	EPVE6AP / EP**E**V6AP	-	-	100*	-
8	MGC16153 / MGC16**5**13	-	-	4	-
9	WBSCR5 / WBS**R**C5	-	-	25*	-
10	FLJ16052 /FLJ160**2**5	100	1	1	1
11	ATP6G1 / **TA**P6G1	-	-	1*	-
12	MGC17922 / MC**G**17922	-	-	1*	-
13	FEEL-2 / FE**E**-L2	100	-	2*	-
14	RANBP7 / RANP**B**7	-	-	5	-
15	INrf2 / INf**r**2	100	1	100*	-
16	FLJ38032 / FLJ380**2**3	100	-	10	-
17	PGR21 / P**R**G21	100	-	8	-
18	U25 / U**5**2	100	100	3	1
19	Nip7-1 / Nip-**7**1	100	-	31	-
20	UNQ2492 / UNQ24**9**2	100	-	3*	-
21	CD62P / C**6**D2P	100	-	100	-
22	FPR / **FR**P	100	100	9	5
23	CD32B / CD3**B**2	100	-	100	-
24	RD / **D**R	100	100	100	-
25	ARHGEF14 / ARHGE**F**14	-	-	23*	-
26	LD / **D**L	100	100	100	-
27	RALGPS1A / RALGPS**A**1	-	-	3*	-
28	AUF1B / AUF**B**1	100	-	4*	-
29	FLJ13784 / FLJ137**4**8	3	100	3	-
30	HSRBC / HSR**C**B	100	-	10*	-
	pourcentage de réussite	0%	0%	50%	0%

TAB. 5.5 – **Recherche approximative - réarrangement :** La recherche des symboles incorrectement orthographiés prouvent que la plupart des moteurs de recherches ne permettent pas de retrouver les gènes en question. Si certains outils ont pu retrouvé certains gènes avec la substitution. Aucun à part Ge-Di ne réussit dans les réarrangements. Les chiffres représentent les faux positifs. Le symbole '*' représente la réussite de la recherche et '-' signifie qu'aucun résultat n'a été généré.

sons que notre plate-forme peut aider partiellement à traiter ces deux types de confusions.

Comme pour tous les tests précédents, nous avons voulu lancer un test supplémentaire. Nous avons voulu mesurer la performance de notre outil sur ces deux types d'erreurs sur un ensemble plus large. Nous avons par conséquent extrait aléatoirement de notre base de données 1066 appellations. Nous avons ensuite généré pour chacune de ces dernières deux approximations. En utilisant ces approximations, nous avons essayé de retrouver le gène en question. Dans cette comparaison, nous avons porté une certaine attention au nombre de faux positifs générés pour chaque entrée.

Pour la substitution, 640 gènes ont pu être retrouvés (60.04 %). Le nombre moyen de faux positifs est de 73.78 avec un écart type de 531.00. L'écart type souligne de grands écarts entre le nombre de faux positifs d'une requête à une autre. L'utilisateur arrive à retrouver son gène 60.04 % des cas cependant il doit traiter un nombre excessif de faux positifs. C'est probablement un inconvénient de notre plate-forme. Pour le réarrangement, ce sont à peu près les mêmes chiffres qui sont retrouvés. 59.00 % des gènes sont retrouvés. La moyenne des faux positifs est de 56.68 avec un écart type 538.31. Encore une fois, nous trouvons que le nombre de faux positifs est élevé. L'écart type traduit de grandes différences d'une requête à une autre. Ces résultats confirment notre crainte. Il y a de grandes chances que le gène soit retrouvé. Malheureusement, l'utilisateur se retrouve avec beaucoup de faux positifs. Comme nous l'avons déjà mentionné, les biologistes s'attaquent souvent à des sous-domaines inconnues. Notre outil de suggestion peut l'aider dans sa quête. Nous pensons cependant qu'il ne l'aide pas dans la certitude d'avoir trouvé l'information recherchée. Un certain nombre de collègues critiquent Google pour ce dernier point. Le nombre excessif de résultats fait douter l'utilisateur. *Doit-il consulter les millions de pages pour être sûr d'avoir la bonne information ?* La réponse à cette question est non, les premiers résultats sont en général les plus pertinents. Une autre question survient alors. *Si les premiers résultats sont pertinents, pourquoi est ce que Google génère des millions de résultats ?* La réponse est que Google sait qu'il peut se tromper et que pour certaines requêtes, plusieurs résultats sont requis. Ces deux questions et réponses sont tirées de discussions avec des collègues. Ces questions traduisent peut être un certain idéal. Néanmoins, ces personnes sont les utilisateurs des outils informatiques et leurs opinions sont importantes. Ces dernières tracent des domaines de recherche pour l'amélioration des moteurs de recherche. Concernant les critiques implicites dans les questions précédents, ces dernières s'appliquent également à notre outil. Cependant dans notre cas, la première question est fatale. Et dans notre cas, la réponse est malheureusement oui. Notre outil ne possède pas un mécanisme de tri des résultats. Par conséquent, le vrai positif peut apparaître un peu partout dans les résultats. Nous discuterons dans le chapitre suivant de la

manière d'améliorer notre outil de suggestion.

5.3 Interface de programmation

Dans cette partie, nous présentons une interface de programmation qui dépend de la plate-forme et qui permet à l'utilisateur d'accéder à la base de données à partir d'applications locales.

5.4 Fonctionnement

En construisant la plate-forme, nous avons été conscients que les besoins d'un utilisateur ne pouvaient pas être satisfaits seulement à travers l'interface graphique. Les domaines touchant aux gènes sont trop nombreux pour que nous puissions répondre de manière appropriée à tous les utilisateurs. De plus, certaines expériences en biologie produisent des milliers résultats. Le traitement subséquent de ces résultats peut nécessiter l'extraction de milliers d'informations. Cette tâche peut se faire de manière manuelle cependant ce processus est assez long. Pour éviter cette perte de temps, nous avons construit une interface de programmation[8]. Cette dernière permet à l'utilisateur d'accèder à la plate-forme à partir d'applications locales. Le fonctionnement de l'API est simple :

1. L'utilisateur exécute une requête http[9] avec les variables appropriées (voir Tableau 5.6)
2. Le serveur traite la demande et retourne les résultats en format XML ou texte.

L'interface retourne deux types de résultats. Si l'entrée n'est pas ambiguë (Voir section 5.2.2), l'interface retourne simplement l'information du gène correspondant. Si l'entrée est ambiguë, l'interface retourne une liste d'identifiants de Ge-Di. Cette liste représente tous les gènes correspondants à l'entrée. Ainsi, l'interface de programmation informe l'utilisateur de l'ambiguïté du terme recherché. L'utilisateur

[8]Application Programming Interface ou API

[9]http ://www-lbit.iro.umontreal.ca/GD/search_engine/data/presearch2.php

peut alors lancer des requêtes additionnelles pour chacun des identifiants. Il peut également utiliser l'information contenu dans l'identifiant du gène (Voir section 4.4).

L'interface de programmation reprend également les différentes options du moteur de recherche basé sur Aho-Corasick. Cependant, une différence subsiste dans la flexibilité. Si le moteur de recherche principal permet des erreurs dans la requête, l'API ne le permet pas vraiment. Nous considérons que les entrées de l'interface de programmation sont moins susceptibles de contenir des erreurs. Si elles contiennent des erreurs, l'échec de la recherche est un moyen d'informer l'utilisateur de la mauvaise orthographe d'une des entrées.

Dans la sections suivante, nous présentons une autre interface de programmation : *Entrez Programming utilities* [WBB+07].

variable	**valeurs**	**détails**	**condition**
query	Toute chaîne de caractère	Query est la requête de l'utilisateur	non
data_type (d)	"1" ou "symbols" "2" ou "geneID" "3" ou "hugoID" "4" ou "gediID" "5" ou "mgiID"	Type de la requête	non
format	"4" ou "XML_PROG" "5" ou "TEXT_PROG"	Le format de la sortie	non
search_type	"1" ou "exact" "2" ou "appro"	search_type détermine la précision de la recherche	**d** = "symbols"
strategy	"1" ou "startingwith" "2" ou "endingwith"	strategy détermine la sens de lecture	**d** = "symbols"
species	"11" : humain et souris "10" : souris "01" : humain	species permet des restriction sur l'espèce	**d** = "symbols"

TAB. 5.6 – **Les variables requises pour l'API de Ge-Di :** Pour pouvoir accéder à la plate-forme sans passer par l'interface graphique, l'utilisateur doit exécuter une requête http avec les variables spécifiées ci-dessus. Certaines variables ont des valeurs prédéfinies et d'autres n'en ont pas.

5.5 Entrez Programming Utilities

Entrez Programming Utilities[10] est un outil développé par le National Center for Biotechnology Information de l'institut national de santé des États-Unis. Cette interface permet d'accéder aux différentes bases de données du centre comme Entrez Gene. L'interface de programmation est divisée en plusieurs modules. Chaque module permet d'exécuter une action spécifique sur le serveur. Nous nous sommes intéressés à deux de ces modules : *EFetch*[11] et *ESearch*[12]. Ces deux modules suivent le même fonctionnement de Entrez Gene. Ce dernier génère toujours une liste de résultats même si cette dernière est composée uniquement d'un seul élément. Le fonctionnement des deux modules est le suivant :

1. L'utilisateur soumet une requête http avec les variables appropriés en utilisant ESearch (Voir tableau 5.7)
2. ESearch retourne une liste d'identifiants de base de données[13]
3. L'utilisateur soumet une nouvelle requête http avec les variables appropriés en utilisant cette fois EFetch (Voir 5.8)
4. EFetch retourne l'information pour la liste d'identifiants

Entrez Utilities a défini des procédures pour mieux répondre aux besoins des utilisateurs. Ainsi un utilisateur qui souhaite simplement les identifiants de Entrez Gene, n'a pas besoin de télécharger l'intégralité de l'information des gènes pour chacun des résultats. Pour cela, il utilise simplement ESearch et s'arrête à l'étape 2. Quant à l'utilisateur qui souhaite plus d'information, il n'est pas trop pénalisé non plus. Il doit suivre à la lettre les 4 étapes. Néanmoins, il n'a pas à soumettre les résultats intermédiaires (c'est-à-dire la liste de l'étape 2). ESearch permet la sauvegarde des résultats sur le serveur. L'utilisateur doit simplement activer la variable *usehistory* dans sa première requête. Ainsi, les résultats intermédiaires définissent les valeurs

[10] http ://www.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html
[11] http ://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi ?
[12] http ://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi ?
[13] La base de données est choisie à l'étape 1.

de *WebEnv* et *query_key*. Ce sont ces deux variables ainsi que la variable *db* que l'utilisateur doit soumettre à l'étape 3. Ainsi, Entrez Utilities permet de répondre de manière efficace aux besoins de l'utilisateur.

variable	valeurs	détails
db	"gene" pour Entrez gene	nom de la base de données
usehistory	'y' ou 'n'	sauvegarder les résultats
term	chaîne de caractères	la requête
retmax	entier	limiter le nombre de résultats
retmode	xml ou text	format de la sortie

TAB. 5.7 – **Les variables utilisées pour ESearch** Ce tableau résume les variables que nous avons utilisées pour générer la liste d'identifiants. Nous avons demandé la sauvegarde de la liste en utilisant la variable *usehistory*. Ce choix nous permet d'éviter de soumettre la liste des résultats à EFetch.

variable	valeurs	détails
db	"gene" pour Entrez gene	nom de la base de données
WebEnv	chaîne de caractères	permet d'identifier la requête sur le serveur
query_key	entier	permet d'identifier la requête sur le serveur
retmode	asn ou xml ou text	format de la sortie

TAB. 5.8 – **Les variables utilisées pour EFetch** Ce tableau résume les variables que nous avons utilisées pour accéder à l'information des gènes. Les valeurs WebEnv et query_key sont données dans les résultats de ESearch.

Maintenant que la présentation des deux interfaces est terminée, nous pouvons débuter la comparaison entre les deux interfaces de programmation.

5.6 Extraction des ARNm

5.6.1 Test

Nous avons vu dans l'introduction nos difficultés lors de la recherche des relations d'expressions dans la littérature. Dans cette section, nous tentons de voir si notre API permet d'automatiser l'étape 3 de notre processus de recherche (Voir section 1.2). Nous mesurons par la même occasion la performance de notre interface en la comparant avec celle de Entrez.

Nous avons reconsidéré les 416 entités biologiques impliquées dans les relations d'expression. Ici, Nous préférons utiliser le terme entité car nous suspectons que certaines entités ne soient pas des gènes. Dans ce cas de figure, nous voulons :

1. Infirmer ou confirmer si l'entité biologique est un gène ou pas
2. S'il s'agit d'un gène, extraire les deux identifiants de la séquence de l'ARNm

Pour pouvoir tester les deux interfaces, nous avons construit deux programmes codés en Java. Nous avons utilisé deux machines identiques pour exécuter les deux programmes :
Processeur : AMD Athlon(tm) 64 X2 Dual Core Processor 4400+
Mémoire vive : 4044032 kB
Puisque ces relations sont propres à la souris, nous avons limité l'espace de recherche à la souris dans nos requêtes. À part cette restriction, nous n'avons effectué aucune autre modification qui puisse avantager l'une ou l'autre des interfaces. Ainsi, nous avons désactivé la suggestion dans Ge-Di. Pour Entrez utilities, nous avons considéré uniquement le premier identifiant de la liste pour chaque entrée soumise. De plus, aucune allocation supplémentaire de mémoire n'a été nécessaire lors de l'exécution. Les deux programmes ont été exécutés durant la nuit après 9 heures. Nous avons ainsi suivi les conseils des administrateurs de Entrez Utilities.
Pour l'analyse des résultats, nous avons décidé de vérifier chacun des résultats émis par les programmes. Nous avons tout particulièrement vérifié les résultats lors qu'il y avait une différence entre Ge-Di et Entrez. Un résultat est valide s'il contient l'entrée dans :

- Le symbole officiel
- Le nom officiel
- Les alias

Néanmoins, nous avons accepté quelques résultats qui ne suivaient pas cette règle. Dans ce cas, nous avons vérifié qu'aucun autre gène ne pouvait correspondre à cette appellation. Cette mesure a été appliquée tout particulièrement aux noms composés de gènes.

À la suite de la vérification des résultats, nous avons pu établir un ensemble de référence. Ce dernier regroupe toutes les appellations de gènes qui ont été identifiées manuellement ainsi que ceux identifiées par l'une ou l'autre des API. Cet ensemble nous a permis de mesurer le taux de précision et de rappel des deux API :

précision={{ensemble de gènes}∩{gènes identifiés}}/{gènes identifiés}

rappel={{ensemble de gènes}∩{gènes identifiés}}/{ensemble de gènes}

5.6.2 Résultats et discussion

Le programme de Ge-Di a terminé la tâche en 2 minutes 39 secondes. Durant ce temps, il a identifié 126 entités comme étant potentiellement des gènes. Le programme de Entrez a terminé la même tâche en 49 minutes 17 secondes. Il a pu identifié 310 entités comme étant potentiellement des gènes. Au moment du test, les deux bases de données contiennent à peu près 207,904 appellations de gènes[14]. Après les vérifications de chacun des résultats, nous avons identifié 186 gènes (44.71 % des entités dans la liste). Sur les 126 entités identifiées par Ge-Di, les 126 se sont avérées être des gènes. Sur le 310 entités identifiées par Entrez, seulement 169 correspondent à des gènes dont 35 sont des identifiants de séquences d'ARN. Même si ces identifiants ne représentent pas des noms de gènes valides, leur identification nous permet d'économiser du temps. Cette identification nous permet également de connître le nom du gène correspondant. Pour le taux de précision, Ge-Di se retrouve avec 100.00% tandis que Entrez 54.52 %. Pour le taux de rappel, Ge-Di se retrouve avec 67.74 % et Entrez 90.00 %.

En pratique, la mise en place des deux programmes a été assez simple. Cependant, le format XML de l'API d'Entrez est le même que celui que nous avons critiqué dans la section 4.5. Ce dernier est très ambigu. Sans notre expérience dans le domaine, nous aurions probablement perdu un temps considérable à déterminer les

[14]Nom officiel, symbole officiel et Alias.

balises importantes. Un problème technique est également survenu lors de l'extraction de l'information. Entrez Utilities tout comme Entrez Gene retourne l'information complète du gène. Malheureusement en XML, cette information peut dépasser 8,000,000 de lignes. Pratiquement, la connexion peut interrompre totalement ou momentanément le transfert de cette information. L'interruption s'est traduite par l'arrêt prématuré du programme à deux reprises. Ce phénomène a eu lieu malgré notre respect des recommandations du guide de Entrez Utilitites.

À première vue, les résultats semblent donner l'avantage à Entrez. Ce dernier permet d'identifier plus de gènes. Cependant, une analyse plus approfondie des résultats semble plutôt souligner les différences entre les interfaces et leurs conséquences sur les résultats. L'interface de programmation de Entrez utilise le même moteur de recherche que l'interface graphique. Nous avons souvent parlé dans ce mémoire, de sa flexibilité. Dans ce test, sa flexibilité est en même temps un avantage et un inconvénient. La flexibilité lui fait générer un nombre assez important de faux positifs. Ce nombre est traduit par le taux peu élevé de précision (54.52 %). D'une certaine manière, on peut même affirmer que pour chaque deux résultats, l'un est un faux positif. Mais la flexibilité de Entrez n'est pas qu'un inconvénient. Il lui permet de retrouver certains gènes à partir d'un nom composé incorrectement orthographié. Cette performance se traduit par le taux de rappel de 90 %.

Le même constat peut être fait pour Ge-Di. Certaines personnes pourraient critiquer la précision de Ge-Di. En étant trop précis, il risque de générer des faux négatifs. Effectivement, des faux négatifs sont générés (60 gènes non identifiés). Son taux de rappel est de 67.74 %. Mais un certain avantage découle de la précision de Ge-Di. Aucun travail supplémentaire n'est nécessaire. Les gènes trouvés peuvent être directement traités. Ce n'est pas vraiment le cas pour Entrez. Pour ce dernier, nous avons eu à éliminer presque la moitié (310-169=141) des entités identifiées. Nous pensons que l'avantage d'une automatisation est perdu si nous avons à vérifier et à éliminer la moitié des résultats. Il faut également préciser que le moteur de Entrez n'est pas sensible à la casse. Cette caractéristique peut expliquer le nombre élevé de faux positifs. Il est vrai également que les biologistes ne respectent pas tou-

jours les nomenclatures de HGNC ou MGI (Voir section 1.5). Néanmoins considérer qu'un gène humain est strictement équivalent à un gène de la souris peut être dangereux au niveau biologique. En attendant que les biologistes soient plus strictes dans leurs articles, il vaut mieux informer l'utilisateur de la mauvaise orthographe du terme recherché plutôt que le considérer équivalent. À date, seulement l'interface graphique encourage indirectement l'utilisateur à vérifier l'orthographe du mot à travers les estimateurs de résultats.

Dans la vérification des résultats, nous avons inclus les identifiants de séquences comme des vrais positifs. Cette démarche a grandement favorisé Entrez Utilities. Son moteur par défaut se concentre sur l'ensemble de l'information du gène tandis que Ge-Di se concentre uniquement sur les appellations. Les séquences identifiées par Entrez existent également dans notre base de données. Cependant, nous n'avons mis aucun moyen à la disposition de l'utilisateur pour qu'il puisse les retrouver. C'est peut-être une amélioration à rajouter à la plate-forme.

Nous pensons que ce test semble plutôt tracer une relation de complémentarité entre les deux interfaces de programmation. Cette complémentarité permettrait d'exploiter les forces de chacune des interfaces tout en évitant les faiblesses. Un tel hybride pourrait peut être atteindre un taux de précision de 100% et un taux de rappel de 90 %.

Finalement, nous aurions voulu englober d'autres interfaces de programmation pour la comparaison. Malheureusement, il n'existe pas une multitude de base de données qui proposent des interfaces de programmation. Nous en avons trouvé qu'une et nous l'avons comparée avec Ge-Di. Nous pensions en trouver une autre dans Gene Card (Voir section 5.2.1) : *gene A La Cart*[15]. À notre grande surprise, l'outil ne s'est pas avéré être une interface de programmation. L'outil est plutôt une interface graphique qui permet de soumettre une liste de gènes et d'extraire l'information correspondante. Cependant, l'utilisateur doit s'inscrire avant de pouvoir utiliser l'outil. Une fois cette démarche faite, l'utilisateur peut effectué jusqu'à 100 requêtes par

[15]http ://www.genecards.org/BatchQueries/index.php

jour. Néanmoins, le site mentionne que l'utilisateur peut faire une demande pour un quota plus élevé. Nous n'avons pas tenté de faire la demande mais nous supposons que l'augmentation du quota n'est pas gratuite. De plus, il ne faut pas oublier que pour les relations d'expression, Gene Card n'est pas éligible car il contient uniquement les gènes humains. Dans les sections suivantes, nous présentons d'autres applications de l'interface.

5.7 Le cross-matching

Les microarray (ou puces à ADN) sont des expériences qui produisent un nombre important de résultats. Ces expériences permettent de mesurer le niveau d'expression de plusieurs gènes en même temps (jusqu'à une dizaine de milliers). Cette comparaison permet de déceler des relations d'expressions entre gènes (voir section 1.2) et de servir de base pour de nouvelles hypothèses. Cette analyse reste néanmoins difficile principalement à cause du nombre de gènes comparés. Il faut également préciser que ce type d'expérience s'est développé très rapidement. Les outils informatiques n'ont pas été développés à la même vitesse [BFB+00]. Une expérience de microarray apporte un certain nombre de renseignements sur l'expression des gènes à un stade donné dans un type donné de cellules. Cependant, il est nettement plus intéressant de combiner les résultats de plusieurs expériences de microarray [MR07]. Ce type de manipulation permet de déceler des nouvelles relations entre gènes mais également de vérifier une relation hypothétique existante.
Avant de pouvoir analyser les résultats de la combinaison de plusieurs expériences de microarray, il faut assigner à chaque gène son niveau d'expression dans chacune des expériences. Or il se peut que l'expérience A utilise le nom officiel d'un gène donné et l'expérience B un des alias du même gène. Un outil informatique est donc nécessaire pour :

1. Assigner à chaque gène un identifiant unique (tel que son symbole officiel)
2. Assigner à chaque gène identifié son niveau d'expression dans chacune des expériences

Ces deux tâches peuvent être effectuées à l'aide l'interface de programmation de Ge-Di. Pour cela il suffit de :

1. Pour chacune des expériences, placer les appellations de gènes dans une liste
2. Pour chaque appellation, vérifier si le gène correspondant existe dans Ge-Di
   - Si c'est le cas, remplacer l'appellation par le symbole officiel du gène
   - Si ce n'est pas le cas, rajouter l'appellation du gène à la liste des gènes non trouvés
3. Pour chaque gène identifié, lui assigner son niveau d'expression dans chacune des expériences

Ainsi, l'interface de programmation permet d'automatiser partiellement le processus du cross-matching. Ici, nous parlons d'automatisation partielle car certaines appellations peuvent ne pas être reconnues par Ge-Di (les faux négatifs). Dans ce cas de figure, l'utilisateur est encouragé à utiliser l'interface graphique de Ge-Di[16] ou tout autre outil pour retrouver le gène en question.

Notre processus de prétraitement n'est pas tout à fait complet car nous n'avons pas abordé les homonymes (voir section 1.4). Si les alias sont pris en charge par notre processus, ce n'est pas le cas des homonymes. Pour ces derniers, l'interface ne fait qu'informer l'utilisateur de l'ambiguïté de l'appellation. Cette démarche lui permet de connaître les gènes correspondants à l'appellation donnée. À partir de cette information, il peut utiliser la métadonnée de la puce ou la littérature pour déterminer le gène en question. C'est peut être un des défauts de l'interface de programmation. Mais à titre de comparaison, l'API de Entrez n'indique en aucun cas l'ambiguïté de l'appellation et l'utilisateur risque bien de se retrouver avec un faux positif (voir section précédente).

Nous avons vu que dans cette application, l'interface de programmation contribue partiellement à résoudre les difficultés. Il est évident que les faux positifs et les faux négatifs vont être présents quelque soit l'outil utilisé. Pour contrer les faiblesses de

[16]Il se peut très bien que l'appellation soit simplement mal orthographiée. Dans ce cas, l'interface graphique peut aider à retrouver le gène en question.

notre outil, nous proposons une intervention manuelle. Ce n'est pas un constat d'échec de la plate-forme ou plus précisément de l'interface de programmation. La bioinformatique ne peut pas éliminer toutes les difficultés en biologie mais elle peut contribuer à en éliminer quelques unes. Dans le domaine des microarray, le développement de nouveaux outils informatiques ainsi que des nouveaux protocoles en biologie sont nécessaires pour obtenir de meilleurs résultats [MR07]. Ici, nous avons présenté un moyen pour faciliter la correspondance entre les gènes et leurs expressions dans le cadre de la combinaison de plusieurs expériences de microarray. Organiser l'information à ce niveau permettrait de faciliter l'analyse subséquente par des outils informatiques [Buc99].

CHAPITRE 6

TRAVAUX FUTURS

Dans le chapitre précédent, nous avons présenté certaines applications de la plate-forme. Cette présentation a été accompagnée de comparaisons avec des outils existants. Ces comparaisons nous ont permis d'évaluer les faiblesses et les forces de notre outil. Dans ce chapitre, nous présentons certaines modifications et améliorations qui pourraient être rajoutées à notre outil.

6.1 À court terme

Optimiser l'estimation : Dans la section 3.5, nous avons décrit l'algorithme d'estimation. Ce dernier effectue une exploration d'une partie de l'arbre en temps réel et retourne le nombre de feuilles qui peuvent être atteintes à partir du lieu d'arrêt. Or nous savons que toute exploration d'un graphe engendre une perte de temps. Il serait donc préférable d'éviter l'exploration. Une solution simple existe pour éviter cette dernière. Il suffit de précalculer pour chaque noeud le nombre de feuilles qui peuvent être atteintes à partir de ce dernier. Ainsi, l'algorithme d'estimation ne fera que lire une valeur et le temps d'attente sera diminué.

Renforcer la grammaire de l'identifiant : Dans la section 4.4, nous avons présenté la grammaire de l'identifiant. Nous avons estimé que notre modèle pouvait couvrir jusqu'à 11 millions d'organismes. Cette estimation se base malheureusement sur un simple dénombrement mathématique et ne prend pas en considération que deux organismes donnés puissent avoir le même identifant. Il faudrait donc renforcer la grammaire pour éviter ce cas. Ce renforcement pourrait être l'ajout de caractères spéciaux tels que des chiffres lorsque deux organismes ou plus possèdent le même identifiant.

Unir les deux outils de recherche : L'architecture actuelle de la plate-forme suppose que l'utilisateur connaisse soit l'appellation soit l'emplacement du gène. S'il possède les deux informations, il est obligé néanmoins d'opter pour l'un des deux outils de recherche et ensuite vérifier si le gène trouvé respecte la deuxième information. Ce cheminement coûteux pourrait être éliminé en créant un module unificateur entre les deux outils. De plus, nous avons vu dans les sections 3.6.4[1] et 5.2.3[2] que les deux outils peuvent générer un nombre excessif de résultats. Par conséquent, ce module permettrait de diminuer le nombre de résultats en réduisant l'espace de recherche.

Améliorer l'expérience de la recherche : Dans le paragraphe précédent, nous avons mentionné que Chromosome Browser était un outil qui pouvait potentiellement générer un grand nombre de résultats. Actuellement, cet outil fonctionne de la manière suivante :

1. L'utilisateur soumet une requête
2. Le serveur traite cette dernière
3. Le serveur retourne les résultats

Or nous avons mentionné dans la section 3.5 que ce fonctionnement n'offre par un contrôle direct sur les résultats. Et que d'autres outils commencent à proposer de nouvelles méthodes plus intuitives. Il serait par conséquent intéressant de proposer certaines de ces méthodes. Nous pensons surtout à l'estimation. Cette dernière permet de voir directement le changement dans les résultats lorsque la requête est modifiée. Il ne faut pas également oublier que certaines requêtes peuvent être des intervalles. Par conséquent, retrouver l'intervalle adéquat nécessite forcement plusieurs modifications de l'entrée. L'estimation serait donc un moyen d'éviter le lancement de requêtes infructueuses sur le serveur.

---

[1] Chromosome Browser

[2] Outil de suggestion du moteur de recherche principal

Etoffer le contenu du dictionnaire : Nous avons dans la section 4.6.2 que HGNC possède 10,229 alias propres. Certains de ces alias sont nouveaux et finiront par être rajoutés dans notre dictionnaire à travers la mise à jour. Pour d'autres, nous devons mettre en place un outil pour les rajouter à notre dictionnaire. Nous devons par la même occasion déterminer si ces ajouts doivent être réguliers. Il suffit peut être d'un seul ajout pour que Ge-Di contienne tous les alias propres de HGNC. Il serait également intéressant de pouvoir compléter les informations disponibles du dictionnaire. Ce dernier actuellement ne propose pas un accès direct aux séquences génomiques. Il faut utiliser le lien vers Entrez Gene à partir de la page des résultats pour les avoir. Rajouter cette information est assez simple. Il suffit d'extraire cette dernière du fichier XML de Entrez Gene (Voir section 4.3). Les séquences génomiques ne sont pas les seules informations à rajouter. Les biologistes s'intéressent de plus en plus aux gènes homologues. Ces sont des gènes qui partagent certaines similarités dans leur séquences et par conséquent peuvent avoir des fonctions similaires. Néanmoins rajouter ce type d'information est un peu plus compliqué. NCBI possède une base de données pour les gènes homologues : HomoloGene[3] [WBB+07]. Son contenu peut être également téléchargé cependant nous ignorons l'organisation des fichiers XML dans cette base de données. Un certain travail doit donc être effectué avant de tenter une extraction. Il faut dire que la prise en main du fichier XML de Entrez Gene n'a pas été aussi simple.

Diversifier les types d'entrées : La plate-forme permet de rechercher un gène à partir d'une de ses appellations ou de son emplacement chromosomique. Or il se peut très bien que l'utilisateur souhaite rechercher le gène correspondant à une séquence donnée. Notre plate-forme n'offre pas cette possibilité actuellement. Ce choix a pénalisé notre outil dans la section 5.6. Nous pourrions rajouter l'option de rechercher un gène en utilisant l'identifiant de sa séquence. Cette option pourrait être rajoutée au moteur de recherche principal. Cependant, certaines procédures

[3]http ://www.ncbi.nlm.nih.gov/entrez/query.fcgi ?db=homologene

doivent être mises en place pour gérer l'espace[4] de recherche et conserver un temps de réponse acceptable.

6.2 À moyen terme

Retrouver un gène à partir d'un sous-mot : L'algorithme d'Aho-Corasick permet de retrouver un gène à partir de son suffixe ou préfixe. Il serait aussi utile de pouvoir retrouver un gène à partir d'un sous-mot c'est à dire toute partie de son appellation qui n'est ni un préfixe ni un suffixe. Cette recherche pourrait être une modification de l'algorithme d'Aho-Corasick tout comme le rajout de la fonction d'échec ou un module indépendant. Nous pensons que la fonction d'échec entraînerait une augmentation de la complexité et que l'algorithme de suggestion devrait être totalement repensé. Nous considérons, par conséquent, qu'un module indépendant d'Aho-Corasick serait plus efficace.

Améliorer la flexibilité de Chromosome Browser : Nous avons discuté dans la section 3.6.2 que Chromosome Browser permettait une certaine flexibilité chez l'humain mais pas chez la souris. Nous avons expliqué notre choix par l'existence de deux types d'emplacements dans cette espèce. Il serait intéressant de faire une table de correspondance entre les deux types d'emplacements. Le problème de conversion a été étudié et une méthode de conversion est même proposée dans la littérature [VMISF04]. Nous pourrions donc utiliser cette dernière pour construire la table. Une fois, cette table faite, nous pourrions offrir la possibilité d'intervalle pour la recherche de gènes de la souris.

Régler le problème des gènes en plus : Le dictionnaire Ge-Di contient 5,654 entrées obsolètes selon Entrez Gene (Voir figure 4.6.2). Étant donné que la littérature peut mentionner ces derniers, nous avons choisi de ne pas les exclure.

[4]En supposant que tout gène possède une séquence génomique et une séquence d'ARNm messager, l'espace de recherche par défaut va passer à 500,000 entrées à peu près.

Cependant, un certain ménage devrait être fait. Avant de nettoyer, il faudrait comprendre les raisons du statut de ces gènes chez Entrez Gene. Nous pourrions contacter les administrateurs du NCBI dans ce cas de figure. Nous pourrions également lancer des analyses supplémentaires sur ces entrées. Nous pensons qu'une entrée obsolète tombe dans deux catégories :

- Entrée remplacée : l'information de l'entrée a été placée dans une entrée
- Entrée supprimée : le gène correspondant à l'entrée n'existe pas (c'est-à-dire que c'est un faux positif dans une expérience donnée)

Les entrées du premier type sont nécessairement à supprimer car il y a redondance d'information. Nous devons cependant conserver les identifiants de Entrez et de Ge-Di. Ainsi l'utilisateur pourrait utiliser ces dernières pour retrouver le gène en question. Offrir cette option n'est pas vraiment possible actuellement. La base de données suppose qu'il existe un seul identifiant de Entrez par gène. Il faudrait alors modifier la table archive pour qu'elle sauvegarde également les anciens identifiants de Entrez Gene. Pour les entrées du deuxième type, ces derniers ne doivent pas être supprimés. Ils ne représentent pas une redondance. Nous devons cependant les marquer avec un statut spécial. Le champ *status* existe déjà dans la base de données. Il se trouve dans la table main (Voir figure 4.1).

6.3 À long terme

Extraction d'information dans le domaine biomédical : Le but principal de créer un dictionnaire de gènes était de pouvoir nous aider dans l'identification des gènes dans la littérature et sur Internet. Nous avons vu dans la section 1.4 qu'il existait une certaine ambiguïté entre les appellations de gènes, les termes de langue anglaise et ceux du vocabulaire biomédical. Nous avons prouvé que notre outil pouvait aider à réduire l'ambiguïté (Sections 5.2.2 et 5.7). Il serait intéressant maintenant de développer des applications d'extraction d'information à partir de Ge-Di. Ce n'est pas uniquement pour pouvoir évaluer ces algorithmes dans le domaine biomédical. Il s'agit plutôt de tirer partie de l'information déjà présente

dans les publications. Ces futurs outils doivent nécessairement inclure des méthodes d'extraction d'information. À partir de la combinaison des connaissances en bioinformatique et dans le domaine de l'extraction d'information, nous allons peut être voir l'apparition de *scientifiques in-silico* [Wre04]. Cependant, il reste un certain chemin avant d'arriver à ce stade. Actuellement, il existe des applications moins nobles mais tout aussi importantes :

- Détection de noms de gènes dans un texte donné[5]
- Recherche d'articles pertinents sur un gène donné
- Résumé automatique d'articles dans le domaine biomédical
- Formulation d'hypothèses en biologie
- Détection d'alias pour les gènes

Les applications sont classées par ordre croissant de complexité.

[5]La plupart des outils ne font pas la distinction entre les gènes ou les protéines comme [CSA04] [TW02] [TXT+05] entre autres.

CHAPITRE 7

CONCLUSION

Dans ce mémoire, nous avons présenté notre expérience dans le développement d'un système pour faciliter la recherche de l'information des gènes sur Internet et dans la littérature. En nous basant sur nos difficultés lors de la recherche des relations d'expression, nous avons construit une plate-forme. Nous avons introduit dans cette dernière un certain nombre de nouveautés :

- Une précision et une flexibilité adaptées aux appellations de gènes
- Estimation exacte du nombre de résultats en temps réel
- Suggestion d'appellations similaires de gènes
- Recherche de gènes à partir d'emplacements chromosomiques
- Élaboration d'une mise à jour spéciale pour la cohérence et la conservation des données

Nous avons également rajouté à la plate-forme une interface de programmation. Ces différents modules ont permis à notre outil d'avoir de bonnes performances par rapport aux systèmes existants. La précision a permis de retrouver l'information des gènes avec une moyenne de 0.09 faux positifs. La flexibilité quant à elle, a retrouvé entre 50 % à 60 % des gènes dont les appellations ont été incorrectement orthographiées. Chromosome Browser a démontré qu'il était plus adapté aux emplacements chromosomiques et qu'il pouvait également être utile dans d'autres applications que sa fonction originelle. Finalement, la mise à jour contribue de manière significative à conserver les données. Cependant, nous pensons qu'un plus grand potentiel existe. Ce dernier peut être achevé en combinant Ge-Di avec d'autres outils tel que Entrez Gene. Cette combinaison permettrait une complémentarité intéressante. Nous retrouvons cette dernière dans les interfaces de programmation où l'une excelle dans la précision et l'autre dans le rappel.

Conscient que notre outil n'est pas parfait, nous avons présenté certaines améliorations pour les défauts existants. Nous avons également tracé certaines perspectives qui

sont à explorer. Dans ce contexte, nous pensons que la plate-forme ouvre plus de portes qu'elle n'en ferme. Nous avons discuté des perspectives à partir de la plate-forme (Section 6.3) mais les idées implémentées dans la plate-forme pourraient être reprises pour d'autres entités dans le monde biologique tels que les protéines ou les ARN.

BIBLIOGRAPHIE

[ABB+03] Victor Ambros, Bonnie Bartel, David P. Bartel, Christopher B. Burge, James C. Carrington, Xuemei Chen, Gideon Dreyfuss, Sean R. Eddy, Sam Griffiths-Jones, Mhairi Marshall, Marjori Matzke, Gary Ruvkun, and Thomas Tuschl. A uniform system for microRNA annotation. *RNA*, 9(3) :277–9, 2003.

[Aho75] Margaret J. Corasick Aho, Alfred V. Efficient string matching : An aid to bibliographic search. *Communications of the ACM*, 18(6) :333–340, 1975.

[BFB+00] T. Beissbarth, K. Fellenberg, B. Brors, R. Arribas-Prat, J. Boer, N. C. Hauser, M. Scheideler, J.D. Hoheisel, G. Schutz, A. Poustka, and M. Vingron. Processing and quality control of DNA array hybridization data. *Bioinformatics*, 16(11) :1014–1022, Nov 2000. Comparative Study.

[Buc99] P. Bucher. Regulatory elements and expression profiles. *Curr Opin Struct Biol*, 9(3) :400–407, Jun 1999.

[Car06] David F. Carr. How Google Works., Jul 2006. [Online; accessed 02-Avril-2007].

[CLF05] Lifeng Chen, Hongfang Liu, and Carol Friedman. Gene name ambiguity of eukaryotic nomenclatures. *Bioinformatics*, 21(2) :248–56, 2005.

[CSA04] Jeffrey T. Chang, Hinrich Schutze, and Russ B. Altman. GAPSCORE : finding gene and protein names one word at a time. *Bioinformatics*, 20(2) :216–225, Jan 2004. Comparative Study.

[DBG+07] Janos Demeter, Catherine Beauheim, Jeremy Gollub, Tina Hernandez-Boussard, Heng Jin, Donald Maier, John C. Matese, Michael Nitzberg, Farrell Wymore, Zachariah K. Zachariah, Patrick O.

Brown, Gavin Sherlock, and Catherine A. Ball. The Stanford Microarray Database : implementation of new analysis tools and open source release of software. *Nucleic Acids Res*, 35(Database issue) :766–770, Jan 2007.

[DNM00] B. Dorohonceanu and C. G. Nevill-Manning. Accelerating protein classification using suffix trees. *Proc Int Conf Intell Syst Mol Biol*, 8 :128–133, 2000.

[EBK+05] Janan T. Eppig, Carol J. Bult, James A. Kadin, Joel E. Richardson, Judith A. Blake, A. Anagnostopoulos, R. M. Baldarelli, M. Baya, J. S. Beal, S. M. Bello, W. J. Boddy, D. W. Bradt, D. L. Burkart, N. E. Butler, J. Campbell, M. A. Cassell, L. E. Corbani, S. L. Cousins, D. J. Dahmen, H. Dene, A. D. Diehl, H. J. Drabkin, K. S. Frazer, P. Frost, L. H. Glass, C. W. Goldsmith, P. L. Grant, M. Lennon-Pierce, J. Lewis, I. Lu, L. J. Maltais, M. McAndrews-Hill, L. McClellan, D. B. Miers, L. A. Miller, L. Ni, J. E. Ormsby, D. Qi, T. B. K. Reddy, D. J. Reed, B. Richards-Smith, D. R. Shaw, R. Sinclair, C. L. Smith, P. Szauter, M. B. Walker, D. O. Walton, L. L. Washburn, I. T. Witham, and Y. Zhu. The Mouse Genome Database (MGD) : from genes to mice–a community resource for mouse biology. *Nucleic Acids Res*, 33(Database issue) :D471–5, 2005.

[EJG+03] Anton J. Enright, Bino John, Ulrike Gaul, Thomas Tuschl, Chris Sander, and Debora S. Marks. MicroRNA targets in Drosophila. *Genome Biol*, 5(1) :R1, 2003.

[ESB06] Ramon A-A Erhardt, Reinhard Schneider, and Christian Blaschke. Status of text-mining techniques applied to biomedical text. *Drug Discov Today*, 11(7-8) :315–325, Apr 2006.

[FCM+04] P. Andrew Futreal, Lachlan Coin, Mhairi Marshall, Thomas Down, Timothy Hubbard, Richard Wooster, Nazneen Rahman, and Mi-

chael R. Stratton. A census of human cancer genes. *Nat Rev Cancer*, 4(3) :177–183, Mar 2004.

[FLSG+06] Stephen V. Faraone, Jessica Lasky-Su, Stephen J. Glatt, Paul Van Eerdewegh, and Ming T. Tsuang. Early onset bipolar disorder : possible linkage to chromosome 9q34. *Bipolar Disord*, 8(2) :144–151, Apr 2006.

[FZ06] Katrin Fundel and Ralf Zimmer. Gene and protein nomenclature in public databases. *BMC Bioinformatics*, 7 :372, 2006.

[HA98] Y. Hosokawa and A. Arnold. Mechanism of cyclin D1 (CCND1, PRAD1) overexpression in human cancer cells : analysis of allele-specific expression. *Genes Chromosomes Cancer*, 22(1) :66–71, May 1998.

[Hun93] Lawrence Hunter, editor. *Artificial intelligence and molecular biology*. American Association for Artificial Intelligence, Menlo Park, CA, USA, 1993.

[JSB06] L.J. Jensen, J. Saric, and P. Bork. Literature mining for the biologist : from information retrieval to biological discovery. *Nat Rev Genet.*, 7(2) :119–129, Mar 2006.

[KK07] Young-Kook Kim and V. Narry Kim. Processing of intronic microRNAs. *EMBO J*, 26(3) :775–783, Feb 2007. Comparative Study.

[Koe05] J. Koehler. Editorial. *Briefings Bionf*, 6(3) :220–221, 2005.

[Lev65] Vladimir I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Doklady Akademii Nauk SSSR*, 163(4) :845–848, 1965.

[LFA93] R. C. Lee, R. L. Feinbaum, and V. Ambros. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75(5) :843–854, Dec 1993. Comparative Study.

[LG05] Mingyi Liu and Andrei Grigoriev. Fast parsers for Entrez Gene. *Bioinformatics*, 21(14) :3189–3190, Jul 2005.

[LH05] Ulf Leser and Jorg Hakenberg. What makes a gene name? Named entity recognition in the biomedical literature. *Brief Bioinform*, 6(4) :357–369, Dec 2005.

[LMBK+07] Peixiang Li, Sarah Maines-Bandiera, Wen-Lin Kuo, Yinghui Guan, Yu Sun, Mark Hills, Guiqing Huang, Collin C. Collins, Peter C. K. Leung, Joe W. Gray, and Nelly Auersperg. Multiple roles of the candidate oncogene ZNF217 in ovarian epithelial neoplastic progression. *Int J Cancer*, 120(9) :1863–1873, May 2007.

[Lud06] Mark D. Ludman. The use of inappropriate, demeaning, and pejorative terminology in gene nomenclature : a comment on Feingold. *Am J Med Genet A*, 140(13) :1485–6, 2006.

[McC76] Edward M. McCreight. A space-economical suffix tree construction algorithm. *Journal of the ACM*, 23(2) :262–272, 1976.

[MCWA85] B. J. Maurer, L. Carlock, J. Wasmuth, and G. Attardi. Assignment of human dihydrofolate reductase gene to band q23 of chromosome 5 and of related pseudogene psi HD1 to chromosome 3. *Somat Cell Mol Genet*, 11(1) :79–85, Jan 1985. Comparative Study.

[Mil95] George A. Miller. Wordnet : a lexical database for english. *Commun. ACM*, 38(11) :39–41, November 1995.

[MOPT05] Donna Maglott, Jim Ostell, Kim D. Pruitt, and Tatiana Tatusova. Entrez Gene : gene-centered information at NCBI. *Nucleic Acids Res*, 33(Database issue) :D54–8, 2005.

[Mor05] Mike Moran. Why web site search users say there are too many results, 2005. [Online; accessed 12-Avril-2007].

[MR04] Sven Mika and Burkhard Rost. NLProt : extracting protein names and sequences from papers. *Nucleic Acids Res*, 32(Web Server issue) :634–637, Jul 2004.

[MR07] Simone Mocellin and Carlo Riccardo Rossi. Principles of gene microarray data analysis. *Adv Exp Med Biol*, 593 :19–30, 2007.

[NBD+06] Jeyakumar Natarajan, Daniel Berrar, Werner Dubitzky, Catherine Hack, Yonghong Zhang, Catherine DeSesa, James R. Van Brocklyn, and Eric G. Bremer. Text mining of full-text journal articles combined with gene expression analysis reveals a relationship between sphingosine-1-phosphate and invasiveness of a glioblastoma cell line. *BMC Bioinformatics*, 7(NIL) :373, 2006.

[PBMW98] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking : Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.

[RCCPL97] M. Rebhan, V. Chalifa-Caspi, J. Prilusky, and D. Lancet. GeneCards : integrating information about genes, proteins and diseases. *Trends Genet*, 13(4) :163, Apr 1997.

[RGJAB04] Antony Rodriguez, Sam Griffiths-Jones, Jennifer L Ashurst, and Allan Bradley. Identification of mammalian microRNA host genes and transcription units. *Genome Res*, 14(10A) :1902–1910, Oct 2004.

[SAB+79] T.B. Shows, C.A. Alper, D. Bootsma, M. Dorf, T. Douglas, T. Huisman, S. Kit, H.P. Klinger, C. Kozak, P.A. Lalley, D. Lindsley, P.J. McAlpine, J.K. McDougall, P. Meera Khan, M. Meisler, N.E. Morton, J.M. Opitz, C.W. Partridge, R. Payne, T.H. Roderick, P. Rubinstein, F.H. Ruddle, M. Shaw, J.W. Spranger, and K. Weiss. International System for Human Gene Nomenclature. *Cytogenet Cell Genet*, 25 :96–116, 1979.

[Sim05] Tom Simonite. Pokemon blocks gene name. *Nature*, 438(7070) :897, Dec 2005. News.

[SMP05] C. Stoll and V. Martel-Petit. Chromosomal region 13q21q31 and heterochrony of development. *Genet Couns*, 16(4) :371–376, 2005. Case Reports.

[SMWK06] M.J. Schuemie, B. Mons, M. Weeber, and J.A. Kors. Evaluation of techniques for increasing recall in a dictionary approach to gene and

protein name identification. *J Biomed Inform*, Sep 2006. JOURNAL ARTICLE.

[SPIBA03] Parantu K. Shah, Carolina Perez-Iratxeta, Peer Bork, and Miguel A. Andrade. Information extraction from full text scientific articles : where are the keywords? *BMC Bioinformatics*, 4 :20, May 2003. Evaluation Studies.

[SWS+04] M. J. Schuemie, M. Weeber, B. J. A. Schijvenaars, E. M. van Mulligen, C. C. van der Eijk, R. Jelier, B. Mons, and J. A. Kors. Distribution of information in biomedical abstracts and full-text publications. *Bioinformatics*, 20(16) :2597–2604, Nov 2004. Comparative Study.

[TPM+99] L.A. Topfer, A. Parada, D. Menon, H. Noorani, C. Perras, and M. Serra-Prat. Comparison of literature searches on quality and costs for health technology assessment using the MEDLINE and EMBASE databases. *Int J Technol Assess Health Care*, 15(2) :297–303, 1999.

[TV06] Javier Tamames and Alfonso Valencia. The success (or not) of HUGO nomenclature. *Genome Biol*, 7(5) :402, 2006. Letter.

[TW02] Lorraine Tanabe and W. John Wilbur. Tagging gene and protein names in biomedical text. *Bioinformatics*, 18(8) :1124–1132, Aug 2002. Comparative Study.

[TXT+05] Lorraine Tanabe, Natalie Xie, Lynne H Thom, Wayne Matten, and W. John Wilbur. GENETAG : a tagged corpus for gene/protein named entity recognition. *BMC Bioinformatics*, 6 Suppl 1 :S3, 2005.

[Ukk95] E. Ukkonen. On-line construction of suffix trees. *Algorithmica*, 14(3) :249–260, 1995.

[VMISF04] Claudia Voigt, Steffen Moller, Saleh M. Ibrahim, and Pablo Serrano-Fernandez. Non-linear conversion between genetic and physical chromosomal distances. *Bioinformatics*, 20(12) :1966–1967, Aug 2004.

[WBB+07] D.L. Wheeler, T. Barrett, D.A. Benson, S.H. Bryant, K. Canese, V. Chetvernin, D.M. Church, M. DiCuccio, R. Edgar, S. Federhen,

L.Y. Geer, Y. Kapustin, O. Khovayko, D. Landsman, D.J. Lipman, T.L. Madden, D.R. Maglott, J. Ostell, V. Miller, K.D. Pruitt, G.D. Schuler, E. Sequeira, S.T. Sherry, K. Sirotkin, A. Souvorov, G. Starchenko, R.L. Tatusov, T.A. Tatusova, L. Wagner, and E. Yaschenko. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, 35 :D5–12, 2007.

[WBL+02] Hester M. Wain, Elspeth A. Bruford, Ruth C Lovering, Michael J. Lush, Mathew W. Wright, and Sue Povey. Guidelines for human gene nomenclature. *Genomics*, 79(4) :464–70, 2002.

[Wei73] P. Weiner. Linear pattern matching algorithm. *14th Annual IEEE Symposium on Switching and Automata Theory*, pages 1–11, 1973.

[Wik07a] Wikipedia. Gène — Wikipedia, the free encyclopedia, 2007. [Online ; accessed 12-Février-2007].

[Wik07b] Wikipedia. Information retrieval — Wikipedia, the free encyclopedia, 2007. [Online ; accessed 12-Février-2007].

[Wik07c] Wikipedia. Medline — Wikipedia, the free encyclopedia, 2007. [Online ; accessed 12-Février-2007].

[WLDP02] Hester M. Wain, Michael Lush, Fabrice Ducluzeau, and Sue Povey. Genew : the human gene nomenclature database. *Nucleic Acids Res*, 30(1) :169–71, 2002.

[WLTB+02] Robert H. Waterston, Kerstin Lindblad-Toh, Ewan Birney, Jane Rogers, Josep F. Abril, Pankaj Agarwal, Richa Agarwala, Rachel Ainscough, Marina Alexandersson, Peter An, Stylianos E. Antonarakis, John Attwood, Robert Baertsch, Jonathon Bailey, Karen Barlow, Stephan Beck, Eric Berry, Bruce Birren, Toby Bloom, Peer Bork, Marc Botcherby, Nicolas Bray, Michael R. Brent, Daniel G. Brown, Stephen D. Brown, Carol Bult, John Burton, Jonathan Butler, Robert D. Campbell, Piero Carninci, Simon Cawley, Francesca Chiaromonte, Asif T. Chinwalla, Deanna M. Church, Michele Clamp, Chris-

topher Clee, Francis S. Collins, Lisa L. Cook, Richard R. Copley, Alan Coulson, Olivier Couronne, James Cuff, Val Curwen, Tim Cutts, Mark Daly, Robert David, Joy Davies, Kimberly D. Delehaunty, Justin Deri, Emmanouil T. Dermitzakis, Colin Dewey, Nicholas J. Dickens, Mark Diekhans, Sheila Dodge, Inna Dubchak, Diane M. Dunn, Sean R. Eddy, Laura Elnitski, Richard D. Emes, Pallavi Eswara, Eduardo Eyras, Adam Felsenfeld, Ginger A. Fewell, Paul Flicek, Karen Foley, Wayne N. Frankel, Lucinda A. Fulton, Robert S. Fulton, Terrence S. Furey, Diane Gage, Richard A. Gibbs, Gustavo Glusman, Sante Gnerre, Nick Goldman, Leo Goodstadt, Darren Grafham, Tina A. Graves, Eric D. Green, Simon Gregory, Roderic Guigo, Mark Guyer, Ross C. Hardison, David Haussler, Yoshihide Hayashizaki, LaDeana W. Hillier, Angela Hinrichs, Wratko Hlavina, Timothy Holzer, Fan Hsu, Axin Hua, Tim Hubbard, Adrienne Hunt, Ian Jackson, David B. Jaffe, L. Steven Johnson, Matthew Jones, Thomas A. Jones, Ann Joy, Michael Kamal, Elinor K. Karlsson, Donna Karolchik, Arkadiusz Kasprzyk, Jun Kawai, Evan Keibler, Cristyn Kells, W. James Kent, Andrew Kirby, Diana L. Kolbe, Ian Korf, Raju S. Kucherlapati, Edward J. Kulbokas, David Kulp, Tom Landers, J. P. Leger, Steven Leonard, Ivica Letunic, Rosie Levine, Jia Li, Ming Li, Christine Lloyd, Susan Lucas, Bin Ma, Donna R. Maglott, Elaine R. Mardis, Lucy Matthews, Evan Mauceli, John H. Mayer, Megan McCarthy, W. Richard McCombie, Stuart McLaren, Kirsten McLay, John D. McPherson, Jim Meldrim, Beverley Meredith, Jill P. Mesirov, Webb Miller, Tracie L. Miner, Emmanuel Mongin, Kate T. Montgomery, Michael Morgan, Richard Mott, James C. Mullikin, Donna M. Muzny, William E. Nash, Joanne O. Nelson, Michael N. Nhan, Robert Nicol, Zemin Ning, Chad Nusbaum, Michael J O'Connor, Yasushi Okazaki, Karen Oliver, Emma Overton-Larty, Lior Pachter, Genis Parra, Kymberlie H. Pepin, Jane Peterson, Pavel Pevzner, Robert

Plumb, Craig S. Pohl, Alex Poliakov, Tracy C. Ponce, Chris P. Ponting, Simon Potter, Michael Quail, Alexandre Reymond, Bruce A. Roe, Krishna M. Roskin, Edward M. Rubin, Alistair G. Rust, Ralph Santos, Victor Sapojnikov, Brian Schultz, Jorg Schultz, Matthias S. Schwartz, Scott Schwartz, Carol Scott, Steven Seaman, Steve Searle, Ted Sharpe, Andrew Sheridan, Ratna Shownkeen, Sarah Sims, Jonathan B. Singer, Guy Slater, Arian Smit, Douglas R. Smith, Brian Spencer, Arne Stabenau, Nicole Stange-Thomann, Charles Sugnet, Mikita Suyama, Glenn Tesler, Johanna Thompson, David Torrents, Evanne Trevaskis, John Tromp, Catherine Ucla, Abel Ureta-Vidal, Jade P. Vinson, Andrew C. Von Niederhausern, Claire M. Wade, Melanie Wall, Ryan J. Weber, Robert B. Weiss, Michael C. Wendl, Anthony P. West, Kris Wetterstrand, Raymond Wheeler, Simon Whelan, Jamey Wierzbowski, David Willey, Sophie Williams, Richard K. Wilson, Eitan Winter, Kim C. Worley, Dudley Wyman, Shan Yang, Shiaw-Pyng Yang, Evgeny M. Zdobnov, Michael C. Zody, and Eric S. Lander. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915) :520–562, Dec 2002. Comparative Study.

[Wre04] Jonathan D. Wren. The emerging in-silico scientist : how text-based bioinformatics is bridging biology and artificial intelligence. *IEEE Eng Med Biol Mag*, 23(2) :87–93, Mar 2004.

[WSVM+03] Marc Weeber, Bob J. Schijvenaars, Erik M. Van Mulligen, Barend Mons, Rob Jelier, Christian C. Van Der Eijk, and Jan A. Kors. Ambiguity of human gene symbols in LocusLink and MEDLINE : creating an inventory and a disambiguation test collection. *AMIA Annu Symp Proc*, pages 704–708, 2003.

[WZL+05] Xiaowo Wang, Jing Zhang, Fei Li, Jin Gu, Tao He, Xuegong Zhang, and Yanda Li. MicroRNA identification based on sequence and structure alignment. *Bioinformatics*, 21(18) :3610–3614, Sep 2005.

[XFH+07] H. Xu, JW Fan, G. Hripcsak, E.A. Mendonca, M. Markatou, and C. Friedman. Gene symbol disambiguation using knowledge-based profiles. *Bioinformatics*, Feb 2007. JOURNAL ARTICLE.

[YHF+02] Hong Yu, Vasileios Hatzivassiloglou, Carol Friedman, Andrey Rzhetsky, and W John Wilbur. Automatic extraction of gene and protein synonyms from MEDLINE and journal articles. *Proc AMIA Symp*, pages 919–923, 2002. Evaluation Studies.

[ZRK+04] Barry R. Zeeberg, Joseph Riss, David W. Kane, Kimberly J. Bussey, Edward Uchio, W. Marston Linehan, J. Carl Barrett, and John N. Weinstein. Mistaken identifiers : gene name errors can be introduced inadvertently when using Excel in bioinformatics. *BMC Bioinformatics*, 5 :80, Jun 2004.

Annexe I

Lexique

ADN ou Acide déxoyribonucéique : se situe dans le noyau de la cellule chez les végétaux et les animaux. L'ADN est formé de deux brins enroulés en hélice. Chaque brin est constitué de l'enchaînement de nucléotides qui diffèrent par une de leurs molécules, qu'on appelle *base*. Il existe 4 bases différentes : adénine (A), thymine (T), guanine (G) et cytosine (C). Les bases maintiennent ensemble les deux brins en formant des ponts hydrogènes.

ARN ou Acide ribonucéique : est formé d'un seul brin. Chaque brin d'ARN est constitué comme l'ADN de l'enchaînement de nucléotides qui diffèrent par une de leurs bases. Il existe 4 bases différentes dans l'ARN : adénine (A), Uracile (U), guanine (G) et cytosine (C). Il existe plusieurs genres d'ARN dans la cellule dont l'ARN messager ($ARNm$) et l'ARN de transfert ($ARNt$).

ARN polymérase : est une enzyme qui sépare les deux brins de l'ADN et catalyse la formation du transcrit primaire.

Épissage : Chez les eucaryotes, les gènes codants des protéines sont constitués d'une suite d'exons et d'introns alternés. Lors de la transcrition, un transcrit primaire est synthétisé, celui-ci va être épissé pour donner lieu à l'ARNm. L'épissage est un processus qui se produit dans le noyau de la cellule et qui consiste en l'excision des introns et en la ligature des exons. L 'ARNm mature, constitué des exons uniquement, est alors exporté vers le cytoplasme pour être traduit en protéine.

Eucaryotes : Se dit des êtres vivants qui ont des cellules dont le noyau est séparé du cytoplasme par une membrane nucléaire. Cet ensemble comprend la souris et l'humain.

Microarray ou puce à ADN : est une collection de molécules d'ADN fixés sur une surface solide. Cette biotechnologie permet de visualiser les gènes exprimés dans une cellule d'un tissu donné et à un moment donné.

Ribosome : est responsable de synthétiser les protéines en décodant l'information contenue dans l'ARNm. Le ribosome est constitué de deux sous-unités, une plus petite qui " lit " l'ARNm et une plus grosse qui se charge de la synthèse de la protéine correspondante.

Thymine : est une base azotée présente dans les acides désoxyribonucléiques du noyau de la cellule. Lors du copiage de l'ADN, la thymine est remplacée par l'uracile dans le transcrit primaire.

Uracile : est une base présente dans l'acide ribonucléique. Le transcrit primaire qui est une copie d'un des brins de l'ADN ne contient pas de thymines (une des bases de l'ADN) mais il contient à la place des uraciles.

Protéine : est une molécule composée d'une chaîne d'acides aminés (il existe 20 acides aminés). La protéine est produite lors de la traduction d'ARNm par le ribosome dans le cytoplasme.

Appendix II

Ge-Di : The Gene Dictionary Platform

Ge-Di : The Gene Dictionary Platform

François Major[1], Amine Halawani

Institute for Research in Immunology and Cancer
Department of Computer Science and Operations Research
Université de Montréal
PO Box 6128, Downtown station
Montréal QC H3C 3J7
Canada

Abstract

Summary: Searching for gene information in literature has always been a difficult task. The lack of formalism in gene annotation, as well as the weakness of nomenclatures rules partially explain this problem. The Gene Dictionary (Ge-Di) is a user-friendly platform made of a search engine and a dictionary that recognizes and corrects human and mouse genes names. When supplied with a gene name, Ge-Di displays the gene's information (ie alias names, location, mRNA sequence, and so on), or, if the entered name is incorrect, it suggests the closest gene names. Ge-Di can also display the genes that are located on a chromosome in a specific user-determined region. Ge-Di is a web application capable of running under Windows, Unix and Mac OS, and can be linked by programming via URL queries.
Availability: http://www-lbit.iro.umontreal.ca/GD
Contact: Francois.major@UMontreal.CA

[1]to whom correspondence should be addressed

Introduction

Research in bio-informatics as well in biology involves searching for genes in literature and often on the world wide web (WWW). More than ever and with major discoveries and results from high throughput experiments, gene searches are constantly increasing in number and complexity. New discoveries are uncovered and new hypotheses are drawn. Unfortunately, this sudden data generation do not translate into a proportional increase of relevant biological knowledge. The speed at which the discoveries are made and slow progress in data formalization can explain the inaccessibility to pertinent information for both humans and computers. In particular, we do not have yet a unique and formal nomenclature for genes, and nothing indicates that it will be the case in a near future.

In the last decade, attempts to solve the problem has emerged. The *Human Genome Nomenclature Committee* (Eyre *et al.*, 2006) (*HGNC* also known as *HUGO*) and the *Mouse Genome Informatics* (Eppig *et al.*, 2005) (*MGI* or *Mouse Genome Nomenclature Committee- MGNC*) are ongoing examples in human and mouse. Their purpose is to assign *official* name and symbol to each single gene. Despite the great efforts made, confusion is still omnipresent and a major problem is the unpopularity of such initiatives (Tamames *et al.*, 2006). Most authors do not use yet the *official* names, or symbols, suggested by *HGNC* (Chen *et al.*, 2005).

Genes information in articles is an essential step to new discoveries, but, for now, any gene related search (in literature) has a high probability of missing relevant information. First, the query does include all possible gene alias(es)[2]. Second, there is no straightforward way to check if a biological term defines a gene or not. Third, the specified gene in literature cannot be retraced to popular databases.

We introduce **Ge-Di**, for Gene Dictionary, a web-based platform designed to identify a gene by its symbol, name, *Entrez Gene* identifier (Maglott *et al.*, 2005), *HGNC* identifier or *MGI* identifier.

[2]A gene alias is a gene name who was either not yet approved or rejected by *HGNC* or *MGI*

Ge-Di Platform Overview

The Gene Dictionary **Ge-Di** is composed so far of a database (Dictionary), an effective search engine and a tool that search for a gene given a specific location (Figure 1B).

The Ge-Di dictionary is extracted from *Entrez Gene* database and formatted to fit Ge-Di organization. While *Entrez Gene*, *HGNC* and *MGI* IDs tend to be pure numbers with low significance to readers in general (Tamames *et al.*, 2006), efforts were made to make Ge-Di IDs meaningful to human and computers. Consequently Ge-Di IDs are composed of three parts. The first part is made of 5 letters and identifies the organism : *HoSap* for Human and *MuMus* for Mouse. The second part depends on the structural chromosome organization and give the location of the genes : chromosome and arm for the Human and chromosome for the mouse. The third part is a number that differentiates neighboring genes. Human gene *p53* (*TP53*), for example is identified as *HoSap17p_368* in the Ge-Di database. In literature, authors prefers to use *p53* despite the confusion with its two known homonyms in the mouse. The goal of the Ge-Di identifiers is not to replace the initiatives of *HGNC* or *MGI* but rather to serve a reliable gene identifier which can be remembered easily. Ge-Di dictionary contains so far 350,000 keywords (names, symbols and aliases) both from Human and Mouse genomes.

The Ge-Di search engine is the most important part the Ge-Di platform. It is based on the Aho-Corasick Algorithm (Aho *et al.*, 1975), known to be one of the most efficient for string matching. While most search engines will either display excessive results or no results at all, this search engine was designed in a way to manage these special cases. For example, searching for *p53* in *Entrez Gene* with species limitation (ie only Mouse and Human) yields more than 700 results instead of the three expected : TP53, Trp53, Trp53-ps (figure 1D).

The Ge-Di algorithm has the ability to switch the reading direction of the search engine: from left to right or from right to left as sometimes it is easier to remember the suffix of a word rather than its prefix. For example, *'232b10HYACC'* is a long

name who is difficult to remember, but it has a distinctive termination, *'ACC'*. The same rules can be applied for *'CCBE1'* gene that can be found using its prefix *'CC'*.

The algorithm precision can be also modified: exact or approximate. The exact mode considers the input as being a complete word, and tries to match it completely in the dictionary. The approximate mode considers the input as being either a suffix or a prefix and tries to find the word in the dictionary containing the input as a suffix or a prefix. By definition a word is also its own suffix and prefix. As an example, given the user input *'ABC'*, the exact mode will try search for the gene *'ABC'* while approximate mode will search for a gene starting or ending with *'ABC'*. When the user input does not match anything, the search engine automatically suggests the closest spelled genes names (figure 1E).

While the user is typing his input, the number of results corresponding to the input so far entered is given (figure 1C: an instant feedback is given to the user). The goal of this feature is to help the user redefine his input when there is no results at all or too many results to check. Case sensitivity is taken in consideration during this process since capitalization is important for genes symbols. *HGNC* guidelines specify that human symbols must be written in upper-case letters while *MGI* specify that mouse symbols must have their first letter in upper-case. The estimation module helps the user avoid the case where the gene name is correctly spelled but the case sensitivity is not respected.

Most search engines will take either database IDs or gene names as inputs, but sometimes the user does not know this information. When the ID or name is missing , a gene location can be used as input. Furthermore, recent studies in literature has shown that neighboring genes may play an important role in evolution (Semon *et al.*, 2006). Thus, we added *Chromosome Browser* option that take as input a chromosomal location (Figure 1B) and displays all genes on the specified location (figure 1F).

Once the gene is found (By either the search engine or *Chromosome Browser*), Ge-

Di displays : all the IDs (*Entrez Gene*, *HGNC* or *MGI* and *Ge-Di*), *official*[3] name and symbol, location(s), gene alias(es), protein alias(e), RNA sequence(s) (figure 1G). Three formats are available for displaying the gene information : HTML, XML, or TEXT. The user can access the selected gene on *Entrez Gene* and *HGNC* (or *MGI*) website as well as it RNA sequence(s) from *Entrez Nucleotide* (Benson *et al.*, 2006) (figure 1H).

The main interface is not the only way to access Ge-Di database. An API (*Application Programming interface*) was designed to allow the user to access Ge-Di through user made programs. Extracting gene information from a list of gene or cross-matching two gene lists can be done using the API. In the field of text mining, new applications can be derived using the Ge-Di dictionary. Recently, a microarray experience was matched with the results of text mining using a dictionary (Nataraja *et al.*, 2006). The complete information about the API are available on the Ge-Di website.

Example

The research process is composed of many tasks. The experimental part is the most famous task because it is usually shared in a written (ie article) or oral form (ie presentation). If the presentation or the article involves genes, the Ge-Di platform can remove ambiguity by detecting polysemy and helping the researcher choose a more appropriate gene name. For example, *CDK1* gene can be very difficult to retrace to a database due to the fact that it is an alias to more than one gene.

Reading articles and listening to presentation is also an important part of the research process. New hypotheses can be drawn by reading articles. For example, a recent article have shown that *p53* might regulate the mitochondrial respiration (Matoba *et al.*, 2006). Some other gene might be also involved in the regulation specially cancer related genes. To prove this hypothesis, *p53* must be knocked out

[3] *As approved by HGNC or MGI*

and the levels of other mRNA in the cell must be analyzed. By using the Ge-Di platform, the *p53* mRNA can be easily extracted.
Last but not least, the research process will not be complete without conferences and symposiums. They are great occasion to share multiple views about a particular subject. However such event has some drawbacks. It is often very difficult to remember all mentioned genes names. Using the remembered gene suffix, prefix or location, the gene can be found very quickly using the Ge-Di platform.

Conclusion

We presented the Ge-Di platform a useful complement to *Entrez Gene* search engine as well as a direct acces to *Entrez gene*, *HGNC* and *MGI* website. In a near future, the Ge-Di platform will be upgraded by additional modules and organisms. Genes are fundamental in biology, but few is known about them. Obtaining the right information can be improved by text mining, but the current trends in biology limits the application these techniques. We believe that Ge-Di is a step forward in the direction of such applications

bibliography

- Aho, Alfred V., Margaret J. Corasick (1975).Efficient string matching: An aid to bibliographic search. *Communications of the ACM* **18(6)** : 333-940.
- Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. (2006) GenBank. *Nucleic Acids Res.* **34** : D16-20.
- Chen L, Liu H, Friedman C. (2005) Gene name ambiguity of eukaryotic nomenclatures.*Bioinformatics* **21(2)** : 248-56.
- Eppig JT, Bult CJ, Kadin JA, Richardson JE, Blake JA, and the members of the Mouse Genome Database Group. 2005. The Mouse Genome Database (MGD): from genes to mice–a community resource for mouse biology. *Nucleic Acids Res* **33** D471-D475
- Eyre TA, Ducluzeau F, Sneddon TP, Povey S, Bruford EA, Lush MJ.(2006) The HUGO Gene Nomenclature Database, 2006 updates. *Nucleic Acids Res* **34** : D319-21.
- Maglott D, Ostell J, Pruitt KD, Tatusova T. (2005) Entrez Gene: gene-centered infor-

mation at NCBI.*Nucleic Acids Res* **33** D54-8.

- Matoba S, Kang JG, Patino WD, Wragg A, Boehm M, Gavrilova O, Hurley PJ, Bunz F, Hwang PM (2006) p53 regulates mitochondrial respiration. *Science* **312(5780)** : 1650-3.

- Natarajan J, Berrar D,Dubitzky W, Hack C, Zhang Y, DeSesa C, Van Brocklyn JR, Bremer EG (2006) Text mining of full-text journal articles combined with gene expression analysis reveals a relationship between sphingosine-1-phosphate and invasiveness of a glioblastoma cell line.*BMC Bioinformatics* **7**:373

- Semon M, Duret L. (2006) Evolutionary origin and maintenance of coexpressed gene clusters in mammals. *Mol Biol Evol* **23(9)** 1715-23.

- Tamames J, Valencia A, (2006) The success (or not) of HUGO nomenclature.*Genome Biol* **15;7(5)** 402.

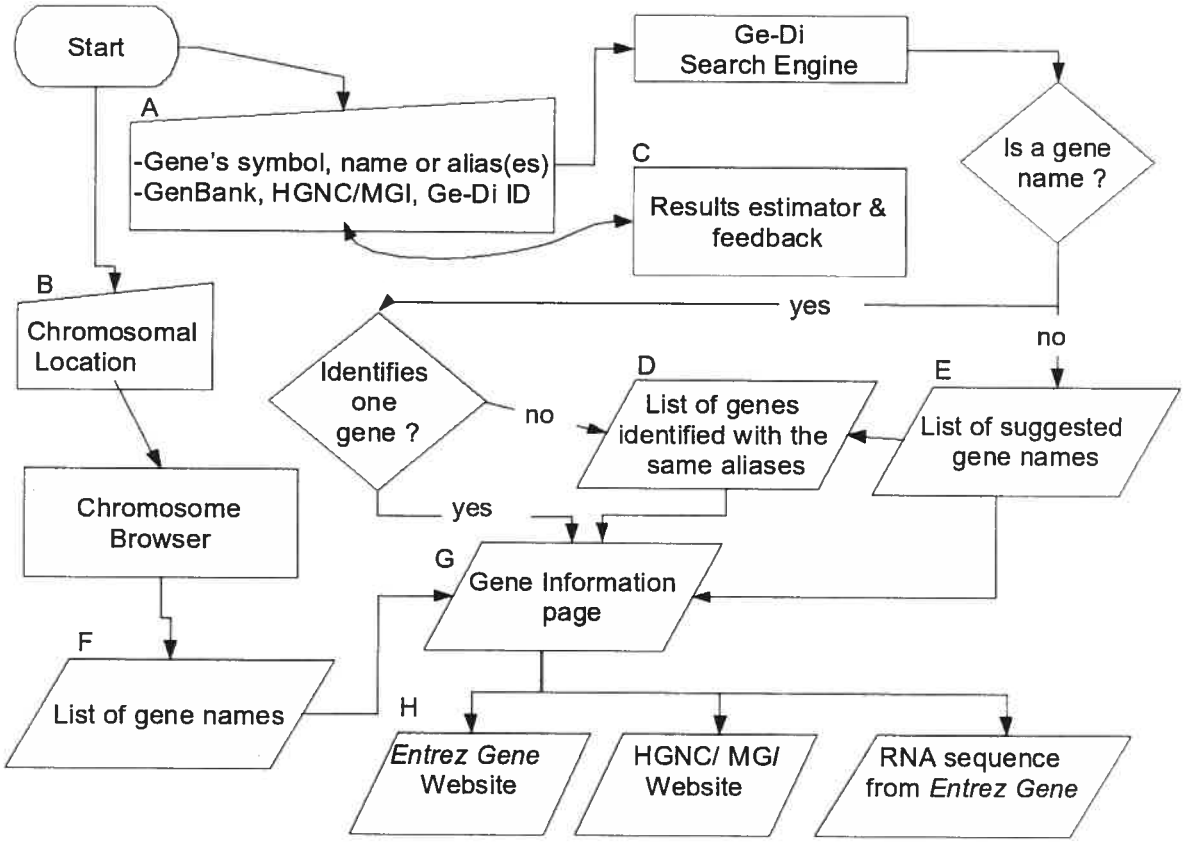

Figure 1: Flowchart of the Ge-Di platform

A trapezoid represents a user input, a rectangle represents a process, a lozenge represents a decision and a parallelogram represents data.

A) represents the kinds of inputs accepted by the Ge-Di search engine.

B) represents the kind of input accepted by *Chromosome Browser*.

C) while the user is typing the query, a result estimator and feedback is displayed on the input page.

D) Sometimes a symbol can be used to identify more than one gene. The user can select the one that corresponds to his interest.

E) Ge-Di always returns a list of suggestions when the approximate search is activated. If the exact search is checked, then Ge-Di will return a list of suggested names if and only if the input does not match any gene's name.

F) Since Browse Chromosome take a user-defined chromosomal location, it always return a list of genes.

G) Once a gene is found, Ge-Di displays its information.

H) From the result page, the user can access directly the selected gene on *Entrez Gene* Website or *HGNC/MGI* website or access its corresponding RNA sequence(s) from *Entrez nucleotide*.