UNIVERSITÉ MONTPELLIER 2
SCIENCES ET TECHNIQUES

CAHIER DES CHARGES

Pré-traitement de données médicales

et

Mise en œuvre de tests

Encadré par

Lisa DI JORIO
Anne LAURENT
Maguelonne TEISSEIRE

Réalisé par

Marc BOURGUÈRE
Guillaume CAMBAZAR
Cleve KENGUE MABIALA
Julien RÉGAL

2008 - 2009

1 Introduction

Notre sujet de TER s'inscrit au sein de deux contextes bien particuliers, puisqu'il s'agit d'appliquer des techniques provenant de la fouille de données à des bases biologiques fournies par l'INSERM. L'objectif de ce TER est donc de coupler les avancées de l'informatique dans le domaine de la fouille de données et d'extraction de connaissance à la recherche médicale menée par l'INSERM. Dans la suite de cette introduction, nous présentons de manière séparée chacun de ces deux contextes.

L' INSERM (Institut National de la Santé et de la Recherche Médicale) est un organisme public français entièrement dédié à la recherche biologique, médicale et en santé des populations. Ses chercheurs ont pour vocation l'étude de toutes les maladies humaines, des plus fréquentes aux plus rares. C'est pourquoi un ensemble de solutions et de techniques d'études plus ou moins poussées sont mises en œuvre. En effet, l'évolution récente des technologies telles que l'acquisition de l'image, ou encore la possibilité d'analyses très poussées mettent à disposition quantités d'informations utiles aux chercheurs de l'INSERM. Ainsi, diverses informations concernant un niveau de granularité très fin (étude des cellules, des gènes, des protéines) viennent s'additionner aux données cliniques déjà disponibles (antécédents familiaux, ages, etc...).

Parmi les maladies traitées se trouve le cancer. Notre projet est en relation directe avec l'étude de cette maladie. Dans ce contexte, les chercheurs de l'INSERM ont réalisé de nombreuses études tumorales, pour lesquelles ils ont réuni diverses informations médicales. D'après les chercheurs de l'INSERM, l'étude d'une multitude de profils permettra la découverte d'informations intéressantes concernant l'évolution de la maladie.

Cependant, la quantité de données à traiter est énorme (une analyse de puce à ADN produit des informations sur des milliers de gènes) et s'avère impossible à traiter manuellement. L'un des objectifs de ce projet est donc de permettre l'automatisation de l'extraction de connaissance à partir de ces données volumineuses. Cet objectif fait partie des thèmes de recherches regroupés sous l'appellation « Extraction de Connaissances dans les grandes bases de Données »(ECD) (ou « Knowledge Discovery in Databases », KDD). Ces thèmes avaient initialement pour but la gestion de données volumineuses mais ils ont évolué avec le temps et visent désormais à tenir compte de l'hétérogénéité des données, de leur format multiple, souvent complexe et de leur qualité variable. La fouille de données est une étape du processus d'ECD et constitue le second contexte de notre TER.

Originellement, les algorithmes d'extraction de données ont été conçus afin d'être appliqués sur des bases où le nombre d'objets (souvent des clients de supermarchés) est largement supérieur au nombre d'items (par exemple des produits en ventes dans un supermarché). Or, l'INSERM produit une très grande quantité de données (items) pour un nombre réduit de patients (objets). Dans de telles circonstances, les méthodes d'extraction classiques arrivent à leurs limites. Il devient alors primordial de proposer de nouvelles méthodes prenant en compte ces diverses spécificités. C'est dans ce contexte que l'équipe TaToo du LIRMM a proposé un algorithme d'extraction de règles graduelles.

Une règle graduelle peut se présenter sous la forme « plus X est A, plus Y est B ». Ce genre de règle appliquée sur des bases de données permet de déterminer des liens entre différents objets d'une même base et d'en extraire une connaissance utile pour la suite. Ce genre de règle appliquée au monde de la biologie, et plus spécifiquement à la recherche sur le cancer, peut permettre aux experts de l'INSERM de repérer des liens entre les différentes données déterminées par leurs recherches.

Ce travail se trouve dans le contexte d'un TER de l'Université Montpellier 2, en collaboration avec l'équipe TaToo du Lirmm. Les thèmes de recherches de cette équipe sont principalement axés sur les processus d'extraction de connaissances, notamment au travers de la conception d'algorithmes d'extraction de règles ou motifs. De manière générale, notre travail consiste en la mise en œuvre de l'extraction de règles graduelles au sein d'un outil facilement exploitable par les experts non informaticiens de l'INSERM.

2 Description des données et des outils

2.1 Puces à ADN

Le cancer est une maladie qui touche la cellule. Celle-ci est constituée d'une membrane et possède en son centre un noyau. Les chercheurs de l'INSERM se concentrent sur ce noyau car c'est ici que se trouvent les instructions nécessaires au développement et au fonctionnement de la cellule. La transformation d'une cellule en cellule cancéreuse est due à l'accumulation d'un nombre suffisant de défauts à l'intérieur du génome d'une cellule. Chez l'homme, le noyau contient 23 paires de chromosomes. Chaque chromosome est composé d'un exemplaire d'origine maternelle et d'un exemplaire d'origine paternelle expliquant la transmission des caractères génétiques de génération en génération. Chaque chromosome contient deux minuscules filaments : c'est la double hélice d'ADN. L'ADN est donc le support de l'information génétique transmise lors du processus de reproduction cellulaire. L'ensemble du matériel génétique d'une espèce est appelé génome. Au-delà de ce rôle de transport d'information génétique, l'ADN sert à la synthèse des protéines, indispensables à la vie. La synthèse des protéines s'effectue en deux grandes étapes : la transcription de l'ADN en ARN messager puis la traduction de celui-ci en protéine. Ce processus est appelé expression de gène.

Nous avons vu qu'une cellule devient cancéreuse suite à l'accumulation suffisante d'un nombre de défauts à l'intérieur de son génome. Cela signifie que les gènes ont subit une modification, donc que le matériel génétique supporté par les gènes a changé, ce qui va également modifier l'expression des gènes. Une manière de comprendre le cancer consiste à étudier ces changements. Pour cela, il est nécessaire de comprendre à quel niveau l'ADN est touché.

Les chromosomes des cellules cancéreuses ont subit d'importantes mutations (quantitative ou qualitative). Ce sont sur ces mutations que les recherches sont focalisées. Les chercheurs utilisent la technique basée sur l'utilisation de puces à ADN pour extraire les différences d'expression de gène entre une cellule cancéreuse et une cellule normale. Une puce à ADN est un ensemble de molécules d'ADN fixées sur une surface qui permet de quantifier le niveau d'expression des gènes dans une cellule d'un tissu donné. Au cours du traitement de ces puces à ADN, il est possible de repérer une perte d'expression de gène, un gain et une expression stable. Les biologistes utilisent la valeur numérique de l'expression de chaque gène, représentée par le ratio gain/perte, seuillé par les valeurs 0.5 et 2. Le logarithme de base 2 est utilisé afin de normaliser les données en vue de tests statistiques. Ces données numériques, couplées aux données cliniques, constituent la carte d'identité d'une tumeur.

2.2 Règles graduelles

Une règle graduelle permet de traduire une variation des valeurs d'éléments entre ensembles (« Plus un camion est lourd, moins il est rapide»). Ces règles ont été majoritairement utilisées dans les systèmes experts fonctionnant à base de règle. Une règle graduelle s'applique sur les valeurs des items et les comparaisons de ces valeurs se font entre les objets.

Considérons la base suivante :

Objet	Age (A)	Salaire (S)
o1	22	1200
o3	24	1200
o2	28	1850
o4	35	2200
o5	38	2000
o6	44	3400
o7	52	3400
o8	41	5000

Un item graduel est de la forme i* avec * ∈ {+,-} (+ signifie « la valeur augmente », - signifie « la valeur diminue ») avec dom(i) muni d'une relation d'ordre totale .

Considérons l'item graduel A^+ (« l'âge augmente ») sur la base précédente. L'objectif de cet item graduel est de sortir des listes d'objets qui répondent à cette contrainte. On remarque que A(o1)< A(o3) < A(o2) < A(o4) < A(o5) où $A(o_i)$ correspond à l'age de l'item o_i. Suite à o5, nous avons plusieurs possibilités : nous pouvons choisir l'objet o6, empêchant à o8 d'apparaître dans la liste, où choisir o8, empêchant o6 d'apparaître. Cette notion de choix implique que plusieurs listes d'objets peuvent respecter un même item graduel. Ainsi, la base composée des objets {o1, o3, o2, o4, o5, o6, o7} et la base composée des objets {o1, o3, o2, o4, o5, o8, o7} répondent à la contrainte imposée par A^+.

Un itemset graduel est un ensemble non vide d'items graduels. A partir des items graduels A^+ et S^+, on obtient l'itemset graduel A^+ S^+. Cet itemset graduel peut être traduit comme : « plus l'âge augmente, plus le salaire augmente », et un ensemble d'objet de la base respectant cette gradualité peut être {o1, o3, o2, o5, o6, o7}. Tout comme l'item graduel, il est important de noter que plusieurs ensembles d'objets peuvent respecter un itemset graduel.

Le support (ou fréquence) d'un itemset graduel reflète le nombre d'objets qui respectent la variation décrite par l'itemset par rapport au nombre d'objets total de la base. La fréquence d'un itemset graduel est calculée à partir de l'ensemble ayant le plus d'éléments. Ainsi, le support de l'itemset graduel A^+ S^+ est de 6/8, soit 75%. Plus généralement, le calcul de cette fréquence est le suivant :

soit s un itemset graduel et G_s l'ensemble des objets respectant s. La fréquence de s est donnée par :

$$Freq(s) = \frac{max|G_s^i|}{|O|} \text{ où } G_s^i \subseteq G_s \text{ et } O \text{ est l'ensemble des objets décrivant la base de données}$$

2.3 De l'utilité des treillis

Les treillis sont utilisés pour représenter des relations d'ordre. Soit un item(set) graduel basé sur un ensemble E. Si la relation d'ordre lui étant associée est une relation binaire réflexive, transitive, antisymétrique et si pour tout (x,y) ∈ E² nous avons x<=y ou x>=y, alors la relation d'ordre liée à l'item(set) graduel est une relation d'ordre total . Cette relation sera notée R_g.

On définit plus généralement cette relation comme suit :

Soit o et o' deux objets d'une base de données BD définie sur un ensemble d'items I={i_1 ... i_n} et R_g une relation d'ordre total sur BD alors $\forall j \in [1,n] \left(o[i_j] R_g o'[i_j]\right) ou \left(o[i_j] R_g o'[i_j]\right)$. On note oR_go'.

Une chaîne est un ensemble formé des éléments a_1, ..., a_n ∈ E tels que $a_i R a_{i+1}$, pour tout i∈[1...n-1].

Les objets d'une base de données qui respectent des itemsets graduels peuvent être représentés à l'aide de treillis. Le calcul de fréquence d'un itemset graduel peut ainsi être ramené au calcul de la longueur de la chaîne composée du plus grand nombre d'éléments dans le treillis représentatif. On divisera cette longueur par le nombre d'objets de la base. Un itemset est dit fréquent lorsque son support dépasse un seuil minimal défini par l'utilisateur. L'extraction d'itemsets graduels consiste donc en la recherche de l'ensemble des itemsets graduels fréquents à partir d'une base de données. Les approches d'extraction de connaissances sont soumises au problème de l'explosion combinatoire et les itemsets graduels ne dérogent pas à la règle. C'est pourquoi l'utilisation des itemsets graduels inverses permet de réduire l'espace de recherche.

2.4 Les itemsets inverses

« Plus l'âge augmente, plus le salaire augmente» et « plus l'âge diminue, plus le salaire diminue » sont deux itemsets inverses. On voit bien qu'à partir de l'un, nous pouvons retrouver l'autre et inversement. Une des propriétés que nous utiliserons concerne l'ensemble des chaînes composant un itemset s et son inverse s' (noté c(s)) : ce sont les mêmes et par extension, Freq(s)=Freq(c(s)). Ceci permet de ne générer que la moitié des itemsets, ce qui permet de réduire l'espace de recherche.

2.5 Algorithme GRITE (Gradual ITemset Extractor)

Les treillis permettent d'organiser les objets en respectant les variations décrites par un itemset graduel. La jointure de deux treillis doit conserver les objets communs entre chaque treillis ainsi que les ordres communs. Cela revient à calculer toutes les sous-séquences communes aux deux treillis. L'opération de jointure étant effectuée un nombre exponentiel de fois, ils faut utiliser une structure de modélisation adaptée : une représentation binaire matricielle de taille n*n pour un treillis de taille n. Les sous-séquences communes aux matrices binaires sont celles dont les bits sont à 1. Ceci est réalisé par l'opération binaire ET entre chaque élément de la matrice. De cette matrice, nous pouvons définir un nouveau treillis représentatif de la jointure d'un itemset graduel s et de son inverse s'. Désormais, il faut déterminer la chaîne de longueur maximale. Cependant, le calcul du plus long chemin est un problème qui peut s'avérer difficile selon les contraintes, mais notre treillis peut se ramener à un graphe orienté et acyclique. Les algorithmes de recherche de plus court chemin son polynomiaux donc il faut poser la contrainte suivante : chaque sommet ne sera considéré qu'une seule fois.

Un sommet peut avoir plusieurs niveaux. En effet, plusieurs ensembles d'objets peuvent respecter un même itemset. Le but étant de maximiser les niveaux, un système de « mémoire » a été mis en place dans l'algorithme, conservant les données obtenues à partir des nœuds de niveau supérieur. Lorsque plusieurs solutions sont possibles, le niveau le plus élevé sera conservé.

L'algorithme, qui sera détaillé lors du rapport final, prend en entrée un nœud et la mémoire. Il donne en sortie la mémoire complétée par les niveaux les plus élevés de chaque nœud. Par conséquent, il suffit de récupérer la valeur maximale de la mémoire pour obtenir la longueur de la chaîne maximale.

3 Travail à réaliser

3.1 Objectif

Notre objectif est de proposer un outil intuitif et convivial pour un utilisateur non informaticien. Au-delà de son interface, nous aurons à écrire divers scripts de transformation des données (pré-traitement) et d'extraction de connaissance (post-traitement) à partir de ces mêmes données. Cette connaissance n'étant pas exploitable directement, nous essaierons de la présenter le plus clairement possible.

Une étude de l'existant devra être faite en amont du travail à réaliser. Nous devons comprendre et intégrer les besoins primordiaux des experts ainsi que leurs attentes face à ce produit. De plus, en comprenant, par divers rendez-vous avec ces mêmes-experts, l'utilisation finale qui sera faite de notre projet, nous parviendrons davantage à nous représenter le produit dans sa finalité.

3.2 Besoins

Pour répondre à ces attentes, nous avons besoin d'acquérir certaines notions. En effet, venant du monde informatique et étant tous novices dans le domaine de la biologie, nous nous devons de comprendre autant que possible le domaine dans lequel nous allons évoluer. Nous avons à nous cultiver pour comprendre la terminologie utilisée et ainsi perdre le moins de temps possible à assimiler ce qui nous est demandé pour passer à l'essentiel le plus rapidement. Au-delà de ce domaine que nous devons appréhender, il nous faudra également comprendre dans ses grandes lignes, puis plus en profondeur, le travail de l'équipe TaToo pour nous permettre de l'exploiter le plus efficacement possible,

3.3 Contraintes

De par l'assimilation du sujet, nous avons soulevé des interrogations et rencontré des problèmes que nous devrons résoudre pour arriver à l'objectif final. Nous devrons donc composer avec les contraintes suivantes (contraintes déterminées jusqu'à présent et donc évolutives dans le temps) :

- les données utilisées par l'INSERM étant des données dites sensibles et confidentielles, nous devrons réaliser nos expérimentations sur des jeux de données générés aléatoirement
- ces mêmes données seront stockées sur des fichiers de type Excel, transformables en fichiers CSV. Plusieurs parseurs devront être écrits afin d'obtenir les informations

nécessaires au travail à réaliser

- la contrainte de compatibilité est primordiale ici. En effet, nous devons réaliser un produit en adéquation avec les contraintes logicielles de l'INSERM. Nous devons y réfléchir plus en profondeur par la suite, mais il semblerait qu'actuellement, notre choix se porte sur le langage PHP car l'INSERM possède un serveur Apache.
- les chercheurs utilisant des machines PC ou Mac, la contrainte de portabilité fait son apparition

3.4 Tâches

Des rendez-vous précédents avec nos tuteurs, nous avons dégagé en leur présence six grandes tâches :

- compréhension du domaine et des règles graduelles
- acquisition des gènes en fonction de leur chromosome et position sur le chromosome
- expérimentations
- pré-traitements et post-traitements des données
- réalisation de l'outil utilisable par l'expert
- tests de l'outil réalisé

La compréhension du domaine et des règles graduelles est une étape primordiale dans la réalisation de notre projet, c'est elle qui conditionnera en grande partie la vitesse d'avancement du projet. En effet, si nous arrivons à maîtriser du mieux possible ce dont il est question et la manière dont nous allons devoir l'implémenter, nous gagnerons un temps important par la suite.

L'acquisition des gènes en fonction de leur chromosome et position sur le chromosome est une partie un petit peu plus délicate. Nous devons associer à chaque gène son nom effectif pour permettre d'être le plus clair possible lors de l'affichage des résultats.

Les tâches d'expérimentation et de traitement de données devront être traitées en quasi simultanées. L'une permettant l'amélioration et la vérification de l'autre.

Puis la réalisation de l'outil utilisable par l'expert sera l'avant-dernière étape de notre projet. Il nous permettra de mettre en application et de regrouper les différentes tâches réalisées auparavant.

Et enfin, divers tests d'utilisation de cet outil seront réalisés pour permettre le « débuggage » et apporter les modifications nécessaires à son bon fonctionnement.

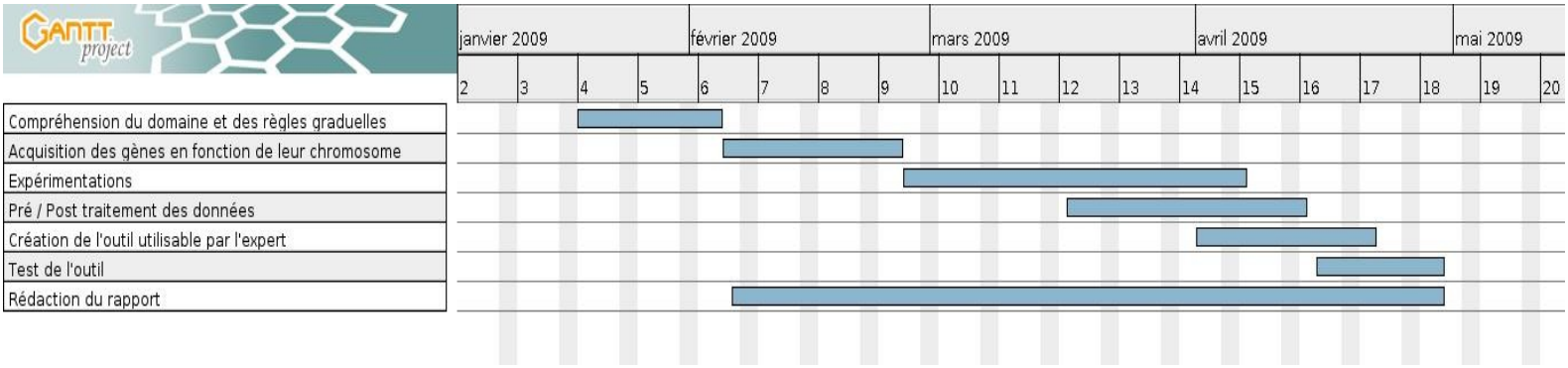
Il ne faut pas oublier que la rédaction du rapport de projet est une tâche existante dans notre planning, qui ne découle pas directement du sujet mais de la réalisation d'un projet en lui-même.

4 Planning

Des sept tâches ci-dessus, nous pouvons déterminer le planning suivant :

- Compréhension du domaine et des règles graduelles : du 19/01/09 au 05/02/09 (3 semaines)
- Acquisition des gènes en fonction de leur chromosome et position sur le chromosome : du 05/02/09 au 26/02/09 (3 semaines)
- Expérimentations : du 26/02/09 au 7/04/09 (5 semaines)
- Pré-traitement / Post-traitement des données du 17/03/09 au 14/04/09 (4 semaines)
- Création de l'outil utilisable par l'expert : du 01/04/09 au 22/04/09 (3 semaines)
- Tests de l'outil : du 15/04/09 au 30/04/09 (2 semaines)
- Rédaction du rapport : du 06/02/09 au 30/04/09 (12 semaines)

Ce planning est modélisé via le diagramme de Gantt suivant :



5 Organisation et communication

5.1 Au sein du groupe

Des séances de travail hebdomadaires seront mises en place en respectant les emplois du temps de chacun afin de faciliter la communication. Des décisions seront ensuite prises à l'issue de ces réunions afin que chaque membre puisse être actif de son côté. La messagerie instantanée, le courrier électronique et le téléphone seront également des moyens de communication utilisés.

5.2 Avec nos tuteurs et les professionnels

Une fois par semaine, nous verrons nos tuteurs afin de leur faire part de l'avancement des travaux, des résultats trouvés et de prendre connaissance de nouvelles directives. Ce moment sera l'occasion également de leur faire part de nos difficultés ou interrogations sur certains points. Nous utiliserons également régulièrement la communication via messagerie électronique, permettant de répondre à de petites interrogations tout en évitant un nouveau rendez-vous.

Il est également prévu que nous rencontrions des chercheurs de l'INSERM. Cette rencontre aura pour but de nous éclairer encore davantage sur le domaine dans lequel nous devons intervenir. En effet, obtenir des informations de personnes dont la formation est axée dans ce domaine sera fortement bénéfique.