Fouille de données biomédicales complexes : extraction de règles et de profils génétiques dans le cadre de l'étude du syndrome métabolique

Sandy Maumus, Amedeo Napoli, Laszlo Szathmary, Sophie Visvikis-Siest

Fouille de données biomédicales complexes : Extraction de règles et de profils génétiques dans le cadre de l'étude du syndrome métabolique

Sandy Maumus[1-2], Amedeo Napoli[2], Laszlo Szathmary[2], Sophie Visvikis-Siest[1]

[1] INSERM U525, 54000 Nancy
{sandy.maumus, sophie.visvikis-siest}@nancy.inserm.fr

[2] LORIA, 54506 Vandoeuvre-Lès-Nancy
{maumus, napoli, szathmar}@loria.fr

Résumé : Nous décrivons une étude sur la fouille de données de la cohorte STANISLAS qui met en œuvre l'extraction de motifs fréquents et de règles d'association. D'une part, nous avons extrait des connaissances utiles et nouvelles pour l'expert qui lui ont permis de générer de nouvelles hypothèses de recherche en biologie. D'autre part, ces travaux nous ont conduit à proposer les premiers éléments d'une méthodologie de fouille de règles extraites en bioinformatique. Au cœur de ces réflexions, nous soulignons le rôle majeur de l'expert dans le processus de fouille de données. Les résultats obtenus sont satisfaisants et illustrent le potentiel des méthodes symboliques de fouille de données en biologie.

Mots-clés : Motif fréquent, Règle d'association, Polymorphisme génétique, Syndrome métabolique.

1 Introduction

Les données recueillies lors des études de populations telles que les études de cohorte sont complexes, d'une part parce que ce genre d'étude se déroule sur une période fixée et les données peuvent donc varier dans le temps, d'autre part parce que ces études intègrent généralement un nombre important d'individus et collectent une quantité de paramètres différents élevée (Mansour-Chemaly *et al.*, 2002). De plus, les données se présentent sous plusieurs types : données quantitatives, qualitatives, textuelle, booléennes, etc... Enfin, le recueil étant fait, il se peut que certaines variables soient manquantes ou leurs valeurs bruitées, ce qui conduit à des bases de données incomplètes. L'ensemble de ces caractéristiques montre que les données issues d'une étude de cohorte peuvent être qualifiées de complexes, que leur volume est important et que leur exploitation est une tâche *a priori* difficile.

En règle générale, les données de ces études sont traitées par des méthodes statistiques, qui constituent la référence pour la majorité des biologistes, épidémiologistes ou médecins confrontés à l'exploitation des résultats. Toutefois, l'avancée technologique en biologie est telle - et notamment en génétique avec le

développement de techniques multiplex et des puces à ADN - que le volume de données à traiter devient de plus en plus grand. Les méthodes statistiques sont robustes mais elles ne suffisent pas à elles seules à exploiter toute la richesse potentielle des données. La problématique principale est ici d'extraire des unités de connaissances à partir de ces données issues d'études de cohorte, qui soient nouvelles et potentiellement utiles, et c'est là que peuvent intervenir les méthodes symboliques de fouille de données. Les techniques symboliques de fouille de données commencent à trouver une place en biologie. Quelques revues biologiques et médicales ont d'ailleurs intégré dans leurs thèmes la bioinformatique et l'extraction de connaissances (Boulicaut & Gandrillon, 2004). Une recherche rapide par mots clés sur PubMed nous a conduit aux résultats rapportés dans la table 1. Une des conclusions qui peut être tirée est que l'application des techniques symboliques de fouille de données, comme la recherche de motifs fréquents et l'extraction de règles d'association, est à l'heure actuelle faiblement utilisée en biologie. Nous nous proposons de promouvoir l'étude et l'utilisation de ces techniques symboliques en fouille de données biologiques issues de cohortes (ou fouille de cohortes).

Pour ce qui nous concerne, nous disposons, dans l'équipe 4 de l'unité INSERM 525, d'une base de données qui contribue à la compréhension des mécanismes entraînant l'athérosclérose : la cohorte STANISLAS (Siest *et al.*, 1998). Un des thèmes auxquels nous nous intéressons plus particulièrement est le syndrome métabolique, une affection prédisposant au diabète de type 2 et aux maladies cardiovasculaires. Le syndrome métabolique (SM) regroupe un ensemble de facteurs de risque cardiovasculaire dont les principaux sont l'insulinorésistance, la dyslipidémie, l'hypertension et l'obésité. Trois définitions ont été proposées pour diagnostiquer le SM chez un individu : une par l'Organisation Mondiale de la Santé (WHO consultation, 1999), une par l'EGIR (Groupe Européen d'Insulino-Résistance) (Balkau & Charles, 1999), et une par le NCEP (National Cholesterol Expert Panel (National Institutes of Health, 2001)) qui sera utilisée par la suite. Ces définitions diffèrent par les critères utilisés et les seuils retenus. Le SM atteint 20 à 25% des individus aux Etats-Unis. En France, le phénomène devient également préoccupant. En outre, nous avons récemment montré que le SM est aussi présent chez les individus de la cohorte STANISLAS, individus pourtant supposés sains : 8,4% des hommes et 6,4% des femmes de la cohorte sont touchés par cette affection (prévalence ajustée sur l'âge et calculée selon les critères du NCEP) (Maumus *et al.*, 2005a). Ces éléments ajoutés à l'augmentation inquiétante du nombre de personnes touchées par l'obésité, y compris les enfants, font comprendre que le SM est devenu dans nos sociétés industrialisées un enjeu majeur de santé publique.

Le but de notre expérimentation est de voir comment les techniques symboliques de fouille de données peuvent nous aider à mieux comprendre les mécanismes physiopathologiques du SM en déterminant les facteurs influençant le SM et avec quelle force. L'application d'une méthode symbolique de fouille de données, en l'occurrence l'extraction de motifs fréquents et de règles d'association, à partir de la cohorte STANISLAS est une expérience originale et nouvelle qui est détaillée dans cet article. Pour l'occasion, un logiciel d'extraction de motifs fréquents et de règles d'association mis au point dans l'équipe Orpailleur a été utilisé (Szathmary & Napoli,

2005). Les résultats de l'expérience viennent corroborer notre idée de compléter les méthodes statistiques usuelles par des méthodes de fouille symboliques.

L'article est organisé de la manière suivante. La section 2 rappelle certaines bases théoriques sur lesquelles repose l'extraction de motifs et de règles et présente les logiciels CORON et ASSRULEX. La section 3 décrit les données de la cohorte STANISLAS. La section 4 présente le détail des expérimentations que nous avons menées sur la cohorte. Enfin, dans la section 5, nous concluons par un bilan des résultats et nos perspectives de travail.

Table 1. Résultats de recherches par mots clés dans PubMed au 01/04/05

Mots clés	**Nombre de documents trouvés**
Data mining	800
Statistics	333840
Itemset search	1
Pattern search	34
Association rules	35
Association rule extraction	12
Lattice	13339
Formal concept analysis	7

2 Les motifs fréquents et les règles d'association : bases théoriques, présentation de CORON et ASSRULEX et état de l'art

Bases théoriques

Ces bases théoriques s'appuient pour une bonne part sur l'article (Bastide et al, 2002) et la thèse (Bastide, 2000).

En entrée, comme pour l'analyse de concepts formels (Ganter & Wille, 1999), nous considérons une relation binaire R entre un ensemble d'objets $O=\{o_1, o_2, ..., o_m\}$ et un ensemble d'attributs $A=\{a_1,a_2,...,a_n\}$. Un contexte binaire est un triplet (O, A, R), où $R \subseteq O \times A$. La notation $R(o,a)$ signifie que l'objet o possède l'attribut a. Le contexte binaire est également appelé *base de données formelle* ou *base de transactions*. Un sous-ensemble d'attributs de A est appelé *motif* et la taille (ou longueur) du motif correspond au nombre d'attributs qui composent le motif. Un motif $P \subseteq A$ est inclus dans un objet $o \in O$ si $(o,p) \in R$ pour tout $p \in P$.

Soit f la fonction qui associe à chaque motif $P \subseteq 2^A$ (où 2^A dénote l'ensemble des parties de A) l'ensemble de tous les objets contenant P: $f(P) = \{o \in O \mid o \text{ contient } P\}$. $f(P)$ est appelée l'image de P.

Le *support* d'un motif indique le nombre d'objets contenant P (le support est éventuellement normalisé par le nombre total d'objets). Un motif est *fréquent* si son support est supérieur à un seuil minimal donné (*support minimum* ou *minsup*). Un motif X est appelé *motif fermé fréquent* s'il n'existe pas de sur-ensemble propre Y (X

$\subset Y$) avec le même support. L'extraction de motifs fréquents consiste à déterminer dans une base de données formelle tous les motifs fréquents ainsi que leurs supports.

Une règle d'association est une implication pondérée mettant en jeu des motifs qui se présente sous la forme suivante: $I_1 \Rightarrow I_2$, où I_1 et I_2 sont des motifs (I_1, $I_2 \in 2^A$) et $I_1 \cap I_2 = \varnothing$. La partie gauche de la règle, I_1, est appelée *antécédent* (ou *prémisse*), tandis que la partie droite, I_2, est appelée *conclusion* (ou *conséquent*). La *confiance* d'une règle d'association r: $I_1 \Rightarrow I_2$ est la probabilité conditionnelle qu'une transaction contienne I_2 sachant qu'elle contient aussi I_1: $conf(r) = supp(I_1 \cup I_2) / supp(I_1)$ où $I_1 \cup I_2$ dénote la concaténation des motifs I_1 et I_2. Le *support* d'une règle d'association r est défini par : $supp(r) = supp(I_1 \cup I_2)$. Une règle d'association est valide si son support et sa confiance sont supérieurs ou égaux aux seuils donnés pour le support minimal (*minsup*) et pour la confiance minimale (*minconf*), respectivement. L'extraction de règles d'association dans une base de données formelle D consiste à découvrir les règles valides dans D.

CORON et ASSRULEX pour la découverte de règles d'association informatives

La découverte de règles d'association est l'une des tâches les plus importantes en fouille de données. La première étape permettant d'accéder aux règles d'association est la détermination des motifs fréquents dans la base de données formelle. Agrawal *et al.* (1996) ont introduit une méthode permettant de générer toutes les règles d'association à partir de tous les motifs fréquents. Cependant, cette méthode produit un nombre colossal de règles d'association et conduit à un nouveau problème de fouille de données que l'on peut désigner par « fouille de règles » : plus l'ensemble des motifs fréquents est grand, plus les règles pouvant être générées sont redondantes[1]. Ainsi l'extraction de règles à partir de motifs fréquents n'est peut-être pas la meilleure solution étant donné le grand nombre de règles produites.

Un autre cadre pour l'extraction de règles d'association est donné par les motifs fermés fréquents. L'ensemble des motifs fermés fréquents est un sous-ensemble des motifs fréquents et il correspond à une représentation complète et condensée des motifs fréquents (Boulicaut & Crémilleux, 2004), ce qui sous-entend que la totalité des motifs fréquents peut être retrouvée à partir des motifs fermés fréquents, avec les supports appropriés (Pasquier *et al.*, 1999). L'ensemble des motifs fermés fréquents est généralement plus restreint que l'ensemble des motifs fréquents, en particulier dans le cas des bases de données formelles denses (les contextes binaires possèdent beaucoup de 1). Un ensemble réduit de règles d'association peut donc être engendré à partir des motifs fermés fréquents. Bastide *et al.* (2002) ont proposé le concept de règles *informatives*. Ces règles ont pour antécédent un élément générateur ou clé, qui est minimal et pour conséquent un fermé, qui est maximal. Un motif clé est minimal (au sens de l'inclusion des motifs) et le motif fermé est maximal dans une classe d'équivalence qui est déterminé par la relation « avoir la même image que » où l'image est donnée par la fonction f introduite plus haut (ensemble des objets qui

[1] Une règle redondante est une règle pouvant être déduite à partir d'une autre règle (Bastide et al, 2002).

possèdent le motif). Tous les motifs d'une même classe ont le même support (puisqu'ils ont la même image). De plus, les règles d'association informatives, qui sont des règles d'association minimales non redondantes, sont les plus intéressantes pour l'utilisateur (Pasquier, 2000).

Dans l'équipe Orpailleur, nous développons une plate-forme appelée CORON pour l'extraction de motifs fréquents et de règles (Szathmary & Napoli, 2005). La version implémentée actuelle propose notamment les algorithmes Apriori, Apriori-Close, Close, Pascal, Titanic et Zart (Bastide *et al.* , 2002 ; Stumme *et al.*, 2002). Grâce à ces algorithmes, il est possible d'extraire les motifs fréquents, les motifs fermés fréquents, ou les deux. Parmi les caractéristiques de CORON figure le fait que chacun des algorithmes cités précédemment peut être appelé en ligne de commande comme un programme indépendant et que CORON peut être facilement intégré dans d'autres systèmes (grâce à son API Java). Les modules qui composent CORON peuvent être réutilisées et assemblés pour implémenter de nouveaux algorithmes de recherche par niveaux. Ainsi, un module appelé ASSRULEX (Association Rule eXtractor), intégré dans CORON, permet la génération de règles d'association à la fois à partir des motifs fréquents et des motifs fermés fréquents.

Dans le travail présenté ici, nous nous sommes plutôt intéressés à la découverte de règles informatives dans la cohorte STANISLAS. L'algorithme Zart (une variation étendue de l'algorithme Close (Pasquier *et al.*, 1999)) fournit la liste des motifs fermés fréquents extraits à partir de la cohorte STANISLAS. En utilisant ces motifs fermés fréquents, ASSRULEX génère les règles d'association informatives. Le nombre des règles d'association informatives est significativement moindre par rapport au nombre total de règles possibles. Ce nombre dépend de la densité de la base de données, mais on estime que le facteur entre les deux ensembles de règles peut varier de 2 à 100.

Travaux connexes

Comme nous l'avons fait remarquer en introduction, les règles d'association sont très peu utilisées en biologie. Quelques études sont apparues ces dernières années (par exemple : Becquet *et al.*, 2002 ; Creighton & Hanash, 2003 ; Quentin-Trautvetter *et al.*, 2002 ; Salleb *et al.*, 2004 ; Stilou *et al.*, 2001), mais rares sont les études travaillant sur des données réelles et d'actualité comme celles de la cohorte STANISLAS. Creighton et Hanash (2003) utilisent pour leur étude des données d'expression de gènes obtenues sur des levures et provenant d'une base de données publique. Ces auteurs utilisent l'algorithme Apriori (Agrawal & Srikant, 1994). De plus, ces auteurs ont choisi de ne s'intéresser qu'aux règles dont la prémisse est de taille 1, avec la possibilité de sélectionner les règles générées à partir de motifs dont l'utilisateur spécifie la taille ou bien qui incluent un item particulier. Leur étude est proche de la nôtre, d'une part du point de vue de la méthode, car Apriori est l'algorithme de recherche par niveau précurseur d'une série de nombreux autres dont fait partie Zart et d'autre part du point de vue de la focalisation sur certains types de

règles. Dans l'étude de Quentin-Trautvetter *et al.* (2002), le logiciel CBA[2] (Classification Based on Association) a été utilisé pour extraire des règles d'association. Confrontés à un nombre de règles généré trop grand pour détecter les règles intéressantes, ces auteurs ont choisi de se restreindre à un nombre plus limité d'attributs et ont alors trouvé des règles intéressantes, ce qui nous conforte dans notre approche qui utilise les projections verticales (sélection d'attributs) et horizontales (sélection d'individus).

3 Application à des données réelles : la cohorte STANISLAS

La cohorte STANISLAS (Suivi Temporaire Annuel Non Invasif de la Santé des Lorrains Assurés Sociaux) est une étude familiale qui a été lancée en 1993 au centre de Médecine Préventive de Vandoeuvre-lès-Nancy. Son objectif principal est d'étudier le rôle et la contribution de facteurs génétiques et environnementaux sur la fonction cardiovasculaire (Siest *et al*., 1998). C'est une étude longitudinale sur dix ans, où des familles de la Meurthe-et-Moselle et des Vosges ont été invitées à venir passer un examen de santé par la Caisse primaire d'assurance maladie tous les cinq ans. Lors du recrutement initial (1993-1995, t_0), les critères d'inclusion étaient les suivants : familles supposées saines, exemptes de maladies aiguës et/ou chroniques, composées de deux parents et de deux enfants de plus de six ans. 1006 familles (4295 sujets) ont ainsi pu être recrutées. Lors de la deuxième visite (1998-2000, t_{+5}), 75% des familles sont revenues. La troisième visite (2003-2005, t_{+10}) est actuellement en cours de réalisation. Les données recueillies dans la cohorte STANISLAS peuvent se diviser en trois catégories : (1) cliniques et environnementales, (2) biologiques et (3) génétiques.

Données cliniques et environnementales

Les données cliniques se divisent en examens cliniques systématiques (mesures morphologiques telles que taille, poids, mesure de la pression artérielle et de la fréquence cardiaque, électrocardiogramme, évaluation du développement pubertaire, cycle féminin…) et en examens cliniques sur projets spécifiques (échographie cardiaque, impédance bioélectrique, ostéodensitométrie…). Par ailleurs, pour ce qui est des données environnementales, des questionnaires sont remplis pour chaque individu et concernent le comportement alimentaire, l'activité physique et professionnelle, la croissance de l'enfant, les antécédents familiaux cardiovasculaires, les habitudes de vie et santé (problèmes de santé et antécédents personnels, consommation d'alcool et de tabac, loisirs, activité professionnelle, diplômes, lieu de résidence), la prise de médicament, la situation de famille et professionnelle.

Données biologiques

Des dosages systématiques sont réalisés : (i) biochimie du sérum (albumine, apolipoprotéines AI et B, aspartate aminotransférase, alanine aminotransférase,

[2] http://www.comp.nus.edu.sg/~dm2/

bilirubine, calcium, cholestérol total, HDL-cholestérol, créatinine, gamma-glutamyltransférase, glucose, phosphatase alcaline, phosphore, protéines, triglycérides, urates) ; (ii) numération de formule sanguine ; (iii) analyse d'urine (recherche de sang, protéines, glucose). Par ailleurs, des dosages spécifiques sur projet sont faits, entre autres : apolipoprotéine (apo) AIV, apo CII, apo CIII, apo E, Lp(a), Lp AI, LpCIIIB, fibrinogène, insuline, vitamines.

Données génétiques

Les données génétiques de chaque individu sont déterminées par la méthode PCR Multiplex[3]. Ainsi, pour chaque sujet, nous disposons de 116 SNPs (Single Nucleotide Polymorphisms ou polymorphismes génétiques[4]) correspondant à tous les processus métaboliques impliqués dans les maladies cardiovasculaires : métabolisme lipidique, pression artérielle, coagulation, adhésion cellulaire et inflammation.

4 Le détail des expérimentations : Une première proposition de méthodologie pour la fouille de règles

Dans les études de cohorte, la population recrutée est plutôt homogène et vérifie de ce fait certains standards qui sont étudiés (notamment pour établir des valeurs de référence pour les paramètres biologiques quand il s'agit de cohortes de sujets sains). Cependant, l'analyste – le spécialiste du domaine qui guide le processus de fouille de données – peut également être intéressé par certains sujets de la cohorte qui possèdent certaines particularités pouvant être porteuses d'informations importantes (par exemple dans notre étude, l'analyste s'intéresse aux individus étant atteints par le SM, un profil rare dans la cohorte STANISLAS où les individus sont des sujets sains). Dans ce cas, plutôt que de parler de motifs fréquents, il faudrait parler de motifs rares. Deux options se présentent alors à l'analyste pour trouver ces motifs rares : baisser le seuil du support minimal ou extraire des motifs non fréquents. L'approche retenue ici par l'analyste est l'utilisation d'un seuil de fréquence très bas.

Dans notre approche, l'implication de l'analyste dans le processus de fouille de données est donc importante et son rôle doit s'étendre du pré-traitement au post-traitement des données. Au niveau du pré-traitement, il doit intervenir en sélectionnant les attributs d'intérêt pour son étude (sélection de colonnes), voire les individus susceptibles d'intérêt (sélection de lignes). Par exemple dans notre expérimentation, en sélectionnant les individus susceptibles d'être atteints par le SM, notre objectif est d'arriver à caractériser le profil génétique associé au SM dans la cohorte STANISLAS. Le processus d'extraction de règles dépend de l'analyste. C'est en effet lui qui précise le support minimal et la confiance minimale. Au niveau du post-traitement, l'analyste doit être capable de repérer les règles intéressantes. Ces différentes considérations nous ont permis de réfléchir à une méthodologie globale pour la fouille de règles qui est esquissée ci-après.

[3] Roche Molecular Systems, Alameda, CA, USA

[4] variations entre individus dans la séquence de gènes.

1) Etapes de pré-traitement

Préparation de la base de données et format d'entrée. Le choix est laissé à l'analyste entre une collection de formats d'entrée : tableau de booléens, tableau d'entiers où chaque individu correspond à une ligne sur laquelle figure une liste de valeurs d'attributs codés par des entiers ou un tableau de booléens où il est possible d'associer un nom à chaque individu et à chaque valeur d'attribut.

Utilisation des filtres de pré-traitement :

Pré-traitement sur les lignes : projection sur les individus permettant de conserver uniquement les objets (individus) possédant un (ou des) attribut(s) spécifié(s) par l'analyste.

Pré-traitement sur les colonnes : projection sur les attributs permettant (i) de sélectionner uniquement certaines colonnes (pour rendre plus efficace le fonctionnement des modules de fouille), ou (ii) de supprimer certaines colonnes

2) Utilisation d'un logiciel de fouille

Nous supposons que l'analyste a choisi une méthode symbolique de fouille de données : recherche de motifs fréquents ou extraction de règles d'association, ou encore classification par treillis ou arbres de décision. L'analyste choisit un support minimal et une confiance minimale pour générer les motifs et/ou les règles d'association. L'analyste peut également faire un choix sur la forme des règles (règles exactes, règles ayant une partie gauche de taille 1…).

Evaluation intermédiaire :

L'analyste peut faire un retour sur les projections et sur le choix des seuils

3) Fouille de règles (étapes de post-traitement)

La fouille de règles peut s'effectuer grâce à l'utilisation de filtres de post-traitement. L'analyste doit avoir la possibilité de sélectionner les règles présentant l'attribut qui l'intéresse en partie gauche, en partie droite ou indifféremment en partie gauche ou droite. Il peut de la même manière supprimer un attribut qui ne l'intéresse pas. Une référence peut être faite à l'article (Creighton & Hanash, 2003) où seules les règles d'association à une prémisse unique sont conservées. Dans le même ordre d'idées, il doit être possible de classer les règles selon leur support, par ordre croissant ou décroissant, de sélectionner les règles dont le support est compris dans un intervalle [a,b] donné, ou de retourner les règles dont le support est $\leq a$ ou $\geq a$, avec a donné.

4) Visualisation des résultats

La visualisation des résultats générés doit s'adapter à la méthode de fouille symbolique choisie.

Dans le cas où l'analyste choisit de travailler sur les motifs rares, il pourra utiliser les treillis de concepts pour la visualisation des résultats (Jay, 2003).

5) Interprétation, validation des résultats et génération de nouvelles hypothèses de travail validées par les statistiques

La dernière étape du processus de fouille est l'interaction avec le statisticien. L'analyste rapporte les nouvelles connaissances extraites grâce aux motifs fermés

fréquents et aux règles d'association informatives et propose de nouvelles hypothèses à tester, par les statistiques par exemple.

Cette méthodologie est conforme au schéma classique de l'extraction de connaissances dans les bases de données (Fayyad et al, 1996).

Application aux données de la cohorte STANISLAS

Pour les expérimentations concernant l'étude du syndrome métabolique avec Zart et ASSRULEX, l'analyste s'est restreint, dans un premier temps, aux données discrètes de la base pour éviter le problème de la discrétisation des données, qui est une des étapes majeures du pré-traitement d'une grande partie des données cliniques et biologiques de la cohorte STANISLAS.

Comme Zart ne traite pas les données manquantes, l'analyste a sélectionné dans la cohorte une population d'adultes pour lesquels il disposait de tous les polymorphismes génétiques (ou SNPs pour Single Nucleotide Polymorphisms). A ces données, ont été ajoutés le sexe et cinq paramètres biologiques qui figurent dans la définition NCEP ATP-III du syndrome métabolique. Les paramètres sont codés de façon booléenne. Chaque polymorphisme génétique est (sauf cas particulier) subdivisé en trois variables qui correspondent aux trois génotypes[5] possibles pour le polymorphisme : (i) l'individu est homozygote[6] pour l'allèle 1, qui est le plus fréquent dans la population (génotype 11), (ii) l'individu est homozygote pour l'allèle 2 qui est l'allèle rare (génotype 22), (iii) l'individu est hétérozygote[7] et a le génotype 12. La table 2 présente le tableau booléen sur lequel a été appliqué Zart. Au final, l'analyste a travaillé sur une base de données de 308 individus et 235 attributs, dont 101 polymorphismes génétiques.

Table 2. Description des données étudiées avec Zart. H : homme ; F : femme ; Gly : hyperglycémie ; HDL : hypoHDLémie, TG : hypertriglycéridémie ; TA : hypertension ; O : obésité ; SM : syndrome métabolique ; PG : polymorphisme génétique

Paramètres / Individus	H	F	Gly	HDL	TG	TA	O	SM	PG_1 =11	PG_1 =12	PG_1 =22	…	PG_{101} =22
i_1	0	1	0	1	1	1	0	0	0	1	0	…	1
i_2	1	0	0	0	1	0	1	0	0	1	0	…	
i_3	1	0	0	1	0	0	0	1	1	0	0	…	0
i_4	0	1	0	0	0	1	0	0	0	1	0	…	0
…	…	…	…	…	…	…	…	…	…	…	…	…	…
i_{308}	0	1	1	0	0	1	0	0	0	1	0	…	0

[5] A un emplacement donné sur un chromosome, un même gène peut exister sous différentes formes appelées allèles. Les différences entre ces allèles d'un même gène portent sur les variations de séquences. Pour un gène donné, la combinaison des deux allèles situés face à face sur les deux chromosomes homologues s'appelle le génotype.

[6] Un individu est homozygote pour un gène lorsqu'il possède deux allèles identiques de ce gène.

[7] Un individu est hétérozygote pour un gène lorsqu'il possède deux allèles différents de ce gène.

Application à l'étude du syndrome métabolique dans la cohorte STANISLAS

Le thème d'étude choisi étant le SM, une projection horizontale (ou sélection d'individus) sur l'attribut SM permet de ne conserver que les individus atteints par le SM selon les critères du NCEP-ATPIII. 9 individus sont retenus. Rappelons qu'ici, l'objectif de l'analyste est la caractérisation du profil génétique associé au SM dans la cohorte STANISLAS. Pour ses expérimentations, l'analyste a fait le choix de ne conserver que les règles exactes (confiance égale à 100%). La fouille sur la totalité des attributs est très difficile. L'ensemble des individus vérifie une règle informative exacte donnée ci-dessous :

{}➔ SM et APOAI_121GG et APOAIV_347ThrThr et APOAIV_360GluGlu et ADRB3_64TrpTrp et NOS3-948AA et ANP_7ValVal et ENaCa_493TrpTrp et FII_20210GG et IL4R_478SerSer et ADRB2_164ThrThr et CCR3_39ProPro et APOB_71ThrIle et LPL_291AsnAsn et FV_506ArgArg et SELE_554LeuLeu (sup=9 ou 100% ; conf=1).

La règle et le motif fermé fréquent correspondant extraits montrent à l'analyste que tous les individus SM de la base ont le génotype APOB_71ThrIle.

Une projection verticale (c'est à dire une sélection d'attributs) est alors réalisée et l'analyste sélectionne le polymorphisme génétique APOB 71Thr/Ile ainsi que les attributs hyperglycémie, hypertriglycéridémie, hypoHDLémie, hypertension, obésité, SM et homme et femme. AssRuleX génère différentes règles d'association informatives. Pour obtenir des règles d'association informatives qui l'intéressent, l'analyste choisit de baisser de façon importante le support. En effet, comme la cohorte STANISLAS décrit une population en bonne santé, les individus atteints par le SM sont peu nombreux et les règles mettant en lumière ces individus sont obtenues en utilisant des seuils de fréquence bas. L'analyste sélectionne alors toutes les règles contenant « SM » dans la partie gauche ou dans la partie droite. 28 règles sont retenues. Une règle intéressante est la suivante :

SM➔APOB_71ThrIle (sup=9 ; conf=1)

Cette règle s'interprète par : être atteint par le SM implique d'avoir le génotype APOB Thr71Ile. Soulignons que cette interprétation est faite sur la base de 9 sujets et que la règle engendrée a une confiance de 100%. Ce chiffre peut paraître faible, mais il ne l'est pas tant que cela en regard du nombre global d'adultes examinés qui est de 308. Cette règle possède un grand intérêt par rapport à la cohorte comme expliqué ci-après.

La dernière étape du processus de fouille est notre cas l'interaction avec le statisticien. Dans notre exemple, il a été testé sur le polymorphisme génétique de l'APOB 71Thr/Ile si la répartition des génotypes contenant l'allèle Ile (allèle 2, encore appelé allèle rare) en fonction de présenter le SM était due au hasard. Ce test a été réalisé sur un échantillon de la cohorte STANISLAS composé de 740 individus. Afin de disposer d'un nombre de sujets atteints par le SM suffisant pour appliquer les méthodes statistiques, le risque d'être SM a été défini selon la définition D3 étendue du NCEP-ATPIII qui englobe certains traitements (Maumus *et al.*, 2005a). La conclusion est que la répartition des génotypes contenant l'allèle Ile est

significativement différente selon que l'individu présente le SM ou non (test χ^2, p=0,03), ce qui suggère qu'une personne possédant l'allèle rare pour le polymorphisme APOB71Thr/Ile est significativement plus fréquemment atteinte par le SM. Cette conclusion est un peu différente de la règle d'association qui avait permis de générer l'hypothèse testée par les statistiques, puisque ici, tous les individus atteints par le SM ne possèdent pas obligatoirement le génotype Thr71Ile, mais il y a significativement plus d'individus atteints qui possèdent l'allèle rare Ile. Cette information présente un caractère original non retrouvé dans la littérature (Maumus *et al.*, 2005b). Dans tous les cas, ce résultat nous incite à continuer et nous montre que notre approche est bien fondée et apporte des résultats nouveaux et tangibles en biologie.

5 Bilan et perspectives

Les motifs fréquents et les règles d'association générés avec Zart et AssRuleX nous ont permis de fouiller les données d'une cohorte et de proposer une méthodologie globale. Ces outils de fouille sont pour nous le moyen d'énoncer de nouvelles hypothèses de travail. Dans ce cas, le processus est guidé par l'analyste, non seulement par ses connaissances *a priori* du domaine, mais aussi par le cheminement des expérimentations qu'il mène.

Pour la suite de notre travail, nous souhaitons enrichir la méthodologie, notamment pour le traitement d'attributs complexes. Il est aussi prévu d'évaluer le résultat de la combinaison de méthodes de classification et d'extraction de motifs fréquents et de règles d'association pour faire émerger des profils d'intérêt dont l'objectif est de contribuer à une meilleure compréhension des mécanismes mis en jeu dans le syndrome métabolique.

Références

AGRAWAL R. & SRIKANT R. (1994). Fast algorithms for mining association rules in large databases. In Proc. Of the 20th VLDB Conf. p. 487-499.

AGRAWAL R., MANNILA H., SRIKANT R., TOIVONEN H. & VERKAMO A.I. (1996). Fast discovery of association rules. In U.M. FAYYAD, G. PIATETSKY-SHAPIRO, P. SMYTH, & R. UTHURUSAMY Eds. *Advances in Knowledge Discovery and Data Mining.* p. 37-57, Menlo Park, CA, AAAI Press/The MIT Press.

BALKAU B. & CHARLES M.A. (1999). Comment on the provisonal report from the WHO consultation. European Group for the Study of Insulin Resistance (EGIR). *Diabet Med*. 16, p. 442-443.

BASTIDE Y., TAOUIL R., PASQUIER N., STUMME G. & LAKHAL L. (2002). Pascal : un algorithme d'extraction des motifs fréquents. *Technique et science informatiques.* 21, p. 65-95.

BASTIDE Y. (2000). Data Mining : algorithmes par niveau, techniques d'implantation et applications. Thèse de l'Université Blaise Pascal, Clermont-Ferrand, N° d'ordre : D.U. 1257, EDSPIC : 226.

BECQUET C., BLACHON S., JEUDY B., BOULICAUT J.F. & GANDRILLON O. (2002). Strong-association-rule mining for large-scale gene-expression data analysis: a case study on human SAGE data. *Genome Biol*. http://genomebiology.com/2002/3/12/research/0067.1.

BOULICAUT JF & GANDRILLON O (coordinateurs) (2004). Informatique pour l'analyse du transcriptome. Traité IC2. Série Informatique et système d'information. Hermès/Lavoisier.

BOULICAUT JF & CREMILLEUX B (coordinateurs) (2004). Extraction de motifs dans des bases de données. RSTI ISI 9(3/4), Hermès/Lavoisier.

CREIGHTON C. & HANASH S. (2003). Mining gene expression databases for association rules. *Bioinformatics*. 19, p.79-86.

FAYYAD UM, PIATETSKY-SHAPIRO G & SMYTH P. From data mining to knowledge discovery: an overview. In: FAYYAD UM, PIATETSKY-SHAPIRO G, SMYTH P, UTHURUSAMY R, editors. Advances in Knowledge Discovery and Data Mining. Menlo Park, California: AAAI Press / MIT Press, 1996:1-36.

GANTER B. & WILLE R. (1999). Formal concept analysis: mathematical foundations. Springer, Berlin/Heidelberg.

JAY N. (2003). Recherche et interprétation de motifs séquentiels fréquents dans une base de données médicales, Mémoire de DEA, Université Henri Poincaré Nancy 1.

MANSOUR-CHEMALY M., HADDY N., SIEST G. & VISVIKIS S. (2002).Family studies: their role in the evaluation of genetic cardiovascular risk factors. *Clin Chem Lab Med*. 40, p. 1085-1096.

MAUMUS S., MARIE B., SIEST, G. & VISVIKIS-SIEST S. (2005a). A Prospective Study on the Prevalence of Metabolic Syndrome (MS) among Healthy French Families. Two Cardiovascular Risk Factors (HDL-C and TNF-α) are revealed in MS Offspring. *Diabetes Care* 28, p.675-682.

MAUMUS S., NAPOLI N., ALBUISSON E. & VISVIKIS-SIEST S. (2005b). STANISLAS Cohort: Detection of Interactions Involving Lipid Genes by Combining Data Mining and Statistics (soumis à *Clinical Chemistry and Laboratory Medicine*).

NATIONAL INSTITUTES OF HEALTH. (2001). Third Report of the National Cholesterol Education Program Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III). *Executive Summary. Bethesda, MD, National Institutes of Health, National Heart, Lung and Blood Institute (NIH publ. no. 01-3670)*.

PASQUIER N., BASTIDE Y., TAOUIL R. & LAKHAL L. (1999). Efficient mining of association rules using closed itemset lattices. *Inf. Syst*, 24, p.25-46.

PASQUIER N. (2000). Mining association rules using formal concept analysis. In *Proc. of the 8th International Conf. on Conceptual Structures (ICCS '00)*, p. 259-264. Shaker-Verlag.

QUENTIN-TRAUTVETTER J., DEVOS P., DUHAMEL A. & BEUSCART R.; QUALIDIAB GROUP. (2002). Assessing association rules and decision trees on analysis of diabetes data from the DiabCare program in France. *Stud Health Technol Inform*. 90, p.557-561.

SALLEB A., TURMEAUX T., VRAIN C. & NORTET C. (2004). Mining quantitative association rules in an atherosclerosis dataset. In Proceedings of the PKDD Discovery Challenge 2004 (co-located with the 6th European Conference on Principles and Practice of Knowledge Discovery in databases), p. 98-103, Pisa, Italy.

SIEST G., VISVIKIS S., HERBETH B., GUEGUEN R., VINCENT-VIRY, M., SASS C. , BEAUD B., LECOMTE E., STEINMETZ J., LOCUTY J. & CHEVRIER P. (1998). Objectives, design and recruitment of a familial and longitudinal cohort for studying gene-environment

interactions in the field of cardiovascular risk: the Stanislas cohort. *Clinical Chemistry and Laboratory Medecine.* 36, p. 35-42.

STILOU S., BAMIDIS P.D., MAGLAVERAS N. & PAPPAS C. (2001). Mining association rules from clinical databases: an intelligent diagnostic process in healthcare. *Medinfo*. 10, p.1399-1403.

STUMME G., TAOUIL R., BASTIDE Y., PASQUIER N. & LAKHAL L. (2002) Computing Iceberg Concept Lattices with Titanic. *J. on Knowledge and Data Engineering (KDE)*. 42, p. 189-222.

SZATHMARY L. & NAPOLI A. (2005). CORON: A Framework for Levelwise Itemset Mining Algorithms. In *Supplementary Proceedings of the Third International Conference on Formal Concept Analysis (ICFCA '05)*, p. 110-113, Lens, France.

WHO consultation (1999): *Definition, diagnosis and classification of diabetes mellitus and its complications. Part I. Diagnosis and classification of diabetes mellitus. World Health Organisation, non-communicable disease surveillance.* Technical report NCS/99-2. W.H.O. Geneva.