Actes de la conférence francophone d'apprentissage automatique CAp 2005, Nice (F),
1-3 juin 2005, Presses Universitaires de Grenoble, pp. 313-328.

Approximation de collections de concepts formels par des bi-ensembles denses et pertinents

Jérémy Besson[1,2], Céline Robardet[3] et Jean-François Boulicaut[1]

[1] INSA Lyon, LIRIS CNRS UMR 5205, F-69621 Villeurbanne cedex, France
http://liris.cnrs.fr

[2] UMR INRA/INSERM 1235, F-69372 Lyon cedex 08, France

[3] INSA Lyon, PRISMA, F-69621 Villeurbanne cedex, France
http://prisma.insa-lyon.fr

Résumé : Le calcul de concepts formels, et plus généralement l'usage des treillis de Galois pour l'extraction de connaissances, a motivé de très nombreuses recherches. Grâce à des progrès algorithmiques récents, ces techniques fournissent des motifs particulièrement intéressants pour l'analyse de grandes matrices codant l'expression de milliers de gènes dans des situations biologiques variées. Dans cet article, nous considérons le contexte réaliste, notamment en biologie, où les concepts formels reflètent des associations trop fortes et donc très sensibles au bruit dans les données. Nous étudions l'extraction de bi-ensembles denses et pertinents pour approximer des collections de concepts formels. Le travail est formalisé dans le cadre de l'extraction de motifs sous contraintes par des algorithmes complets. Plusieurs validations expérimentales confirment la valeur ajoutée de notre approche.

Mots-clés : Découverte de connaissances, extraction de motifs sous contraintes, concepts formels, bioinformatique.

1 Introduction

L'extraction de concepts formels dans des contextes booléens et plus généralement l'usage des treillis de Galois pour l'extraction de connaissances ont motivé de nombreuses recherches. Les contextes booléens, également appelés données transactionnelles[1], se retrouvent dans de nombreuses applications. Ainsi, nous travaillons à l'analyse du transcriptome (étude des mécanismes de régulation des gènes chez un organisme vivant) après codage de propriétés d'expression booléennes pour des (dizaines de) milliers de gènes dans des situations biologiques variées. En effet, des techniques

[1]Des données transactionnelles sont un multi-ensemble d'items. Ce type de données souvent étudié en "data mining", correspond à de (grandes) matrices booléennes où les lignes définissent les transactions et les colonnes représentent les items : la présence d'un item dans une transaction est codée par la valeur vrai.

expérimentales comme celles des puces ADN permettent de quantifier le niveau d'expression des gènes (voir, e.g., la matrice de gauche de la figure 1) et dont on peut dériver des données booléennes d'expression (e.g., la matrice de droite de la figure 1). Cette dernière code le fait que les gènes ont ou pas un fort niveau d'expression (ici une valeur >1.5[2]). Dans de tels contextes booléens, un concept formel, ou rectangle maximal de valeurs 1 (vrai), représente un motif a priori intéressant pour les biologistes : il informe sur une association forte entre un ensemble maximal de gènes qui sont co-exprimés et un ensemble maximal de situations biologiques donnant lieu à cette co-expression. L'extraction de tels motifs fournit alors des collections de modules de transcription potentiels permettant d'accélerer la découverte de nouvelles voies de régulation (Besson *et al.*, 2004b), i.e., l'un des objectifs majeurs de l'analyse du transcriptome.

	Gènes			
	g_1	g_2	g_3	g_4
s_1	1.8	2.3	1.6	2.0
s_2	2.1	2.4	0.3	1.1
s_3	1.1	1.6	0.2	0.1
s_4	0.3	0.3	2.1	1.1
s_5	0.25	0.5	0.5	1.0

	Gènes			
	g_1	g_2	g_3	g_4
s_1	1	1	1	1
s_2	1	1	0	0
s_3	0	1	0	0
s_4	0	0	1	0
s_5	0	0	0	0

FIG. 1 – Matrice d'expression de gènes (gauche) et une matrice booléenne $\mathbf{r}_1$ (droite)

Par définition, les concepts formels sont construits sur des ensembles fermés. En marge des algorithmes de calcul de concepts formels (voir (Fu & Nguifo, 2004) pour une synthèse récente), de nombreux chercheurs ont proposé des algorithmes de calcul d'ensembles fermés dits fréquents qui peuvent désormais s'appliquer à de très grandes matrices booléennes (Pasquier *et al.*, 1999; Pei *et al.*, 2000; Zaki & Hsiao, 2002; Goethals & Zaki, 2003). On peut alors calculer des collections de concepts fréquents au sens de (Stumme *et al.*, 2002) : seuls les concepts dont l'un des ensembles est suffisamment grand sont extraits. En s'intéressant aux dimensions très particulières des matrices d'expression booléennes (peu de lignes et de très nombreuses colonnes), (Rioult *et al.*, 2003) montre qu'il est possible d'utiliser n'importe quel algorithme efficace de calcul d'ensembles fermés fréquents[3] sur la plus petite des deux dimensions et ainsi calculer tous les concepts formels dans des données d'expression typiques. Pour traiter des cas plus difficiles, i.e., lorsqu'aucune des deux dimensions n'est suffisamment petite ou lorsque la densité du contexte (nombre de valeurs 1) est trop importante pour les algorithmes existants, nous avons proposé D-MINER, un algorithme complet d'extraction de concepts formels sous contraintes (Besson *et al.*, 2004a). Il permet d'exploiter efficacement les contraintes monotones sur les deux dimensions des concepts formels (e.g., une taille minimale pour chacun des deux ensembles, une "surface minimale", des contraintes d'inclusion).

Nous avons maintenant des preuves de l'intérêt des concepts formels pour l'analyse

[2]Il s'agit d'un codage naïf mais des approches plus réalistes ont été étudiées (Pensa *et al.*, 2004).

[3]On utilise ici avec un seuil de fréquence nulle.

du transcriptome et la découverte de connaissances biologiques (Besson *et al.*, 2004b; Meugnier *et al.*, 2005).

Cependant, dans un concept formel, on capture une association très forte entre un ensemble de gènes et un ensemble de situations. Intuitivement, un concept n'accepte aucune exception. Si le concept $c_1 = (\{s_1, s_2, s_3\}, \{g_1, g_2, g_3, g_4\})$ est considéré comme traduisant une association réelle et si, dans les données, g_3 ne vérifie plus la propriété booléenne pour s_2, alors on trouvera les deux concepts $(\{s_1, s_2, s_3\}, \{g_1, g_2, g_4\})$ et $(\{s_1, s_3\}, \{g_1, g_2, g_3, g_4\})$ mais pas le concept c_1. En fait, la présence de valeurs "indûment" mises à 0 va faire exploser le nombre de concepts formels à extraire. Notons également que l'on aura des problèmes avec des valeurs codées par 1 alors qu'elles auraient du prendre la valeur 0. Dans ces contextes bruités, non seulement les extractions peuvent devenir impossibles, mais aussi les interprétations des motifs calculés sont très difficiles. En d'autres termes, nous sommes en présence d'une très grande sensibilité au bruit. Or, non seulement les données d'expression numériques sont bruitées du fait de la complexité des techniques de mesure, mais aussi le prétraitement de codage des propriétés booléennes à partir des données numériques peut introduire du bruit.

Dans cet article, nous proposons de travailler avec un nouveau type de motif : des bi-ensembles contenant un nombre borné de 0 par ligne et par colonne, et tel que chaque ligne (resp. colonne) soit suffisamment différente de chaque ligne (resp. colonne) extérieure sur l'ensemble des colonnes (resp. lignes) du bi-ensemble. Nous montrons que ce type de motif, appelé *bi-ensemble dense et pertinent*, est plus robuste au bruit et permet en pratique de concentrer davantage d'information pertinente dans des collections de motifs plus petites.

Dans la section 2 nous présentons quelques travaux connexes. La section 3 formalise notre problème dans le cadre de l'extraction sous contraintes. Dans la section 4, nous décrivons succinctement l'algorithme développé pour l'extraction de tous les bi-ensembles denses et pertinents. La section 5 s'intéresse aux résultats expérimentaux obtenus, notamment dans le cas de données biologiques réelles. Nous montrons que même dans le cas où le calcul de tous les bi-ensembles denses et pertinents est trop difficile, on peut utiliser l'algorithme proposé pour étudier les extensions de certains concepts. Enfin, nous concluons dans la section 6.

2 Travaux connexes

Les récentes techniques de bi-partitionnement tendent à fournir des rectangles plus robustes au bruit mais au moyen de recherches heuristiques (optimisations locales) et surtout sans recouvrement (Dhillon *et al.*, 2003; Robardet, 2002). D'autres approches ont été proposées dans la communauté de l'extraction de motifs sous contraintes. Dans (Yang *et al.*, 2001), les auteurs étendent la définition des ensembles fréquents[4] à des ensembles tolérants au bruit. Ils proposent un algorithme par niveau pour les calculer. Malheureusement, ces motifs ne peuvent pas être extraits facilement car les contraintes qui les définissent ne sont ni anti-monotones ni monotones relativement à l'inclusion

[4]Dans notre contexte, un ensemble fréquent correspond à un ensemble de gènes suffisamment co-exprimés au regard d'un nombre minimal de situations biologiques impliquées.

ensembliste, des propriétés essentielles pour rendre les extractions faisables. Ils utilisent donc un algorithme glouton calculant une solution incomplète. Dans (Seppänen & Mannila, 2004), les auteurs recherchent une contrainte anti-monotone. Ils proposent un algorithme par niveau pour calculer les ensembles qui ont une densité de valeurs 1 supérieure à δ dans au moins σ situations. L'anti-monotonicité est obtenue en exigeant que tous leurs sous-ensembles vérifient également cette contrainte. L'extension de tels ensembles denses à des bi-ensembles est difficile : les correspondances qui associent les gènes aux situations biologiques, et réciproquement, ne sont ni croissantes ni décroissantes. En effet, l'ensemble des situations biologiques associé à un ensemble de gènes n'est pas nécessairement inclus dans celui de ses sur-ensembles. Dans (Gionis *et al.*, 2004), les auteurs calculent des motifs ("geometrical tiles") qui sont des rectangles denses (ayant une densité de valeurs 1 supérieure à un seuil fixé). Pour extraire ces motifs, ils utilisent un algorithme non déterministe d'optimisation locale qui ne garantit pas la qualité globale des motifs extraits. Ils exigent qu'il existe un ordre sur les deux dimensions de la matrice : les rectangles ne sont pas considérés à des permutations près des lignes et/ou des colonnes mais doivent concerner des éléments contigus au regard des ordres considérés. Cette hypothèse n'est clairement pas acceptable dans notre contexte.

Une autre approche importante consiste à étudier de façon systématique la notion de représentation condensée des collections de concepts formels ou de bi-ensembles denses, qu'il s'agisse de représentations exactes ou approximatives. L'objectif est alors de ne représenter, ou mieux de ne calculer, qu'un sous-ensemble des collections tout en pouvant retrouver, plus ou moins exactement mais à un faible coût, l'ensemble de la collection. On peut vouloir, par exemple, rechercher une collection de k motifs qui approxime le mieux des collections complètes (Afrati *et al.*, 2004). L'approche des représentations condensées doit aussi intégrer des approches de "zoom" comme, par exemple, les travaux présentés dans (Ventos *et al.*, 2004) pour construire des treillis de Galois à différents niveaux d'abstraction. Cette méthode utilise une partition sur les objets qui permet de réduire le nombre de motifs extraits. Ils utilisent une partition sur les lignes et ne conservent que les concepts qui sont en "accord" avec cette partition : une situation s appartient à l'extension d'un ensemble G si $\alpha\%$ des objets de la même classe que s satisfont G et que s satisfait aussi G. Nous souhaitons pour notre part avoir une approche duale entre les situations et les gènes où aucune des deux dimensions n'est privilégiée au cours de l'extraction.

3 Définitions

Nous notons $\mathcal{G}$ l'ensemble des gènes et $\mathcal{S}$ l'ensemble des situations biologiques. Le contexte à fouiller est booléen, i.e., la représentation d'une relation $\mathbf{r} \subseteq \mathcal{S} \times \mathcal{G}$. Ces situations peuvent correspondre à des expériences de type puce ADN (voir figure 1).

3.1 Bi-ensembles

Un bi-ensemble (S, G) est un couple d'ensembles de $2^{\mathcal{S}} \times 2^{\mathcal{G}}$. Certains bi-ensembles particuliers peuvent être extraits dans des matrices booléennes comme les 1-rectangles (tous les éléments de S sont en relation avec tous les éléments de G) ou les concepts formels qui sont des 1-rectangles maximaux (en fait, S et G sont des ensembles fermés). Les nombreux travaux sur le calcul d'ensembles d'items (typiquement les ensembles fréquents utilisés pour le calcul de règles d'association (Becquet *et al.*, 2002)) peuvent être considérés comme des calculs de bi-ensembles. On associe à un ensemble de gènes toutes les situations qui le "portent" et l'on a donc un 1-rectangle particulier appelé "itemset". D'une manière duale, on peut définir un motif similaire basé sur un ensemble de situations appelé "objectset".

Nous donnons quelques rappels sur les correspondances de Galois (voir notamment (Wille, 1982)) pour formaliser notre problème.

Définition 1 (Correspondance de Galois)
Soit $\phi : \mathcal{S} \rightarrow \mathcal{G}$ et $\psi : \mathcal{G} \rightarrow \mathcal{S}$ deux opérateurs entre deux ensembles partiellement ordonnés $(\mathcal{S}, \leq_{\mathcal{S}})$ et $(\mathcal{G}, \leq_{\mathcal{G}})$. Ces opérateurs forment une correspondance de Galois si :

$$
\begin{array}{ll}
1 & \forall v, w \in \mathcal{S},\ \textit{si}\ v \leq_{\mathcal{S}} w\ \textit{alors}\ \phi(w) \leq_{\mathcal{G}} \phi(v), \\
2 & \forall i, j \in \mathcal{G},\ \textit{si}\ i \leq_{\mathcal{G}} j\ \textit{alors}\ \psi(j) \leq_{\mathcal{S}} \psi(i), \\
3 & \forall v \in \mathcal{S}, \forall i \in \mathcal{G},\ v \leq_{\mathcal{S}} \psi(\phi(v))\ \textit{et}\ i \leq_{\mathcal{G}} \phi(\psi(i))
\end{array}
$$

où $\leq_{\mathcal{S}}$ et $\leq_{\mathcal{G}}$ sont deux relations de spécialisation respectivement sur $\mathcal{S}$ et $\mathcal{G}$.

Définition 2 (Correspondances ϕ et ψ)
Si $S \subseteq \mathcal{S}$ et $G \subseteq \mathcal{G}$, ϕ et ψ peuvent être définis ainsi : $\phi(S, \mathbf{r}) = \{g \in \mathcal{G} \mid \forall s \in S,\ (s, g) \in \mathbf{r}\}$ et $\psi(G, \mathbf{r}) = \{s \in \mathcal{S} \mid \forall g \in G,\ (s, g) \in \mathbf{r}\}$. ϕ renvoie l'ensemble des gènes qui satisfont la propriété d'expression dans toutes les situations biologiques de S. ψ fournit l'ensemble des situations biologiques pour lesquels on a la propriété d'expression de tous les gènes de G. (ϕ, ψ) forme une correspondance de Galois entre $\mathcal{S}$ et $\mathcal{G}$ munis de l'inclusion ensembliste $\subseteq$ (relation de spécialisation). Nous utilisons les notations classiques $h = \phi \circ \psi$ et $h' = \psi \circ \phi$ pour désigner les opérateurs de fermeture de Galois. Un ensemble $S \subseteq \mathcal{S}$ (resp. $G \subseteq \mathcal{G}$) est dit fermé dans $\mathbf{r}$ ssi $S = h'(S, \mathbf{r})$ (resp. $G = h(G, \mathbf{r})$).

On peut maitenant formaliser les types de motifs précités.

Définition 3 (1-rectangles, ensembles et concepts formels)
Un bi-ensemble (S, G) est un 1-rectangle dans un contexte $\mathbf{r}$ ssi $\forall s \in S$ et $\forall g \in G$, $(s, g) \in \mathbf{r}$. Quand un bi-ensemble n'est pas un 1-rectangle, on dit qu'il contient des valeurs 0. Un bi-ensemble (S, G) est un concept dans $\mathbf{r}$ ssi $S = \psi(G, \mathbf{r})$ et $G = \phi(S, \mathbf{r})$. Ceci est équivalent à $S = h'(S, \mathbf{r})$ et $G = \phi(S, \mathbf{r})$ ou à $G = h(G, \mathbf{r})$ et $S = \psi(G, \mathbf{r})$. Une propriété importante de la correspondance de Galois est que chaque ensemble fermé sur l'une des deux dimensions est associé à un unique ensemble fermé de l'autre dimension.

Exemple 1
$(\{s_1\},\{g_1,g_3\})$ et $(\{s_1,s_2\},\{g_2\})$ sont des 1-rectangles dans $\mathbf{r}_1$ mais ne sont pas des concepts. Un exemple de concept dans $\mathbf{r}_1$ est $(\{s_1,s_2\},\{g_1,g_2\})$. Nous avons $h(\{g_1,g_2\},\mathbf{r}_1)=\{g_1,g_2\}$, $h'(\{s_1,s_2\},\mathbf{r}_1)=\{s_1,s_2\}$, $\phi(\{s_1,s_2\},\mathbf{r}_1)=\{g_1,g_2\}$, et $\psi(\{g_1,g_2\},\mathbf{r}_1)=\{s_1,s_2\}$. On peut associer à l'ensemble de gènes $\{g_1\}$ l'ensemble des situations $\{s_1,s_2\}=\psi(\{g_1\},\mathbf{r}_1)$ et nous pouvons alors parler du 1-rectangle $(\{s_1,s_2\},\{g_1\})$ comme d'un itemset. Notons qu'avec nos définitions, le 1-rectangle $(\{s_1,s_2\},\{g_2\})$ n'est pas un itemset : il faudrait ajouter s_3 à sa première composante.

Nous avons motivé dans l'introduction l'intérêt de travailler avec des bi-ensembles qui soient moins sensibles au bruit que les concepts formels et plus pertinents vis-à-vis des données globales. La faisabilité des extractions dépend de l'existence de contraintes monotones et anti-monotones (voir définition 4) permettant de définir les motifs recherchés. En fait, monotonicité et anti-monotonicité sont des propriétés duales qui sont très bien exploitées pour des extractions complètes de motifs sous contraintes, même en présence de grands espaces de recherche.

Définition 4 (Relation de spécialisation et monotonicité)
La relation de spécialisation $\preceq$ que nous utilisons sur les bi-ensembles de $2^{\mathcal{S}}\times 2^{\mathcal{G}}$ est définie par $(S_1,G_1)\preceq(S_2,G_2)$ ssi $S_1\subseteq S_2$ and $G_1\subseteq G_2$. Une contrainte $\mathcal{C}$ est dite anti-monotone par rapport à $\preceq$ ssi $\forall X,Y\in 2^{\mathcal{S}}\times 2^{\mathcal{G}}$ tels que $X\preceq Y, \mathcal{C}(Y)\Rightarrow\mathcal{C}(X)$. $\mathcal{C}$ est dite monotone par rapport à $\preceq$ ssi $\forall X,Y\in 2^{\mathcal{S}}\times 2^{\mathcal{G}}$ tel que $X\preceq Y, \mathcal{C}(X)\Rightarrow\mathcal{C}(Y)$.

Définition 5 (Exemple de contraintes monotones sur les bi-ensembles)
Contrainte de taille minimale : un bi-ensemble (S,G) satisfait $\mathcal{C}_{ms}(\mathbf{r},\sigma_1,\sigma_2,(S,G))$ ssi $\sharp S\geq\sigma_1$ et $\sharp G\geq\sigma_2$ où $\sharp$ désigne le cardinal d'un ensemble.
Contraintes d'inclusion : un bi-ensemble (S,G) satisfait $\mathcal{C}_{Inclusion}(\mathbf{r},X,Y,(S,G))$ ssi $X\subseteq S$ and $Y\subseteq G$.
Contrainte de surface minimale : un bi-ensemble (S,G) satisfait $\mathcal{C}_{area}(\mathbf{r},\sigma,(S,G))$ ssi $\sharp S\times\sharp G\geq\sigma$.

A la recherche de bi-ensembles denses, nous avons proposé dans (Besson *et al.*, 2005) une première approche visant à calculer des bi-ensembles ayant un nombre borné de valeurs 0. La méthode proposée consistait en un post-traitement de la collection de tous les concepts formels. L'idée était de procèder à une fusion de certains concepts de telle sorte que le nombre de valeurs 0 par ligne et par colonne soit borné. Cette contrainte étant anti-monotone suivant $\preceq$, ce procédé peut être réalisé en adaptant un algorithme d'extraction d'ensembles maximaux. Malheureusement, les motifs ainsi extraits ne sont pas munis d'une correspondance de Galois : le même ensemble de situations biologiques peut être associé à plusieurs ensembles de gènes différents. Nous proposons maintenant d'extraire un nouveau type de motif appelé *bi-ensemble dense et pertinent* muni d'une telle correspondance. Il s'agit de calculer tous les bi-ensembles qui satisfont la conjonction des contraintes introduites ci-dessous.

3.2 Bi-ensembles denses

Le concept de densité peut être envisagé sous deux angles selon que l'on mesure le nombre de 0 par ligne/colonne ou sur l'ensemble du bi-ensemble (densité forte versus faible) et selon que l'on considère ce nombre de manière absolue ou relativement à la taille du bi-ensemble (densité absolue versus relative).

La contrainte de "densité forte absolue" impose une limitation du nombre de 0 par ligne et par colonne, mais, relativement à la taille du bi-ensemble, elle borne aussi supérieurement le nombre de 0 total du bi-ensemble. De plus, lorsque le seuil de densité choisi est petit devant la taille minimale du bi-ensemble, ces bi-ensembles ne contiennent pas de lignes et de colonnes presque vides (avec presque que des 0) contrairement à ce qui peut se produire avec la densité faible.

D'autre part, on peut obtenir un résultat similaire sans devoir pousser de contrainte de taille minimale et en utilisant seulement une contrainte de "densité forte relative" : en fixant la proportion de 0 par ligne et par colonne on ne peut obtenir de ligne ou de colonne pleines de 0.

Ainsi, nous souhaitons extraire des bi-ensembles ayant un nombre maximum α de valeurs 0 et contenant au moins γ fois plus de 1 que de 0 par ligne et par colonne. Cette contrainte est notée $\mathcal{C}_d(\mathbf{r}, \alpha, \gamma, (S, G))$.

3.3 Bi-ensembles pertinents

Nous voulons extraire des bi-ensembles composés de situations biologiques ayant une densité sur les gènes du bi-ensemble supérieure à celle sur les gènes n'appartenant pas au bi-ensemble. Réciproquement, le bi-ensemble doit contenir des gènes dont la densité sur les situations biologiques du bi-ensemble est supérieure à celle des situations biologiques n'appartenant pas au bi-ensemble.

De manière plus formelle, étant donné deux paramètres δ, un bi-ensemble (S, G) est dit *pertinent* ssi

$$\begin{aligned} \max_{s \in S}(\sharp\{g \in G \mid (s,g) \notin \mathbf{r}\}) + \delta &\leq \min_{s \in \mathcal{S} \setminus S}(\sharp\{g \in G \mid (s,g) \notin \mathbf{r}\}) \\ \max_{g \in G}(\sharp\{s \in S \mid (s,g) \notin \mathbf{r}\}) + \delta &\leq \min_{g \in \mathcal{G} \setminus G}(\sharp\{s \in S \mid (s,g) \notin \mathbf{r}\}) \end{aligned}$$

Par la suite, cette contrainte sera désignée par $\mathcal{C}_s(\mathbf{r}, \delta, (S, G))$.

Par construction, plus δ augmentent, plus la différence entre la densité du bi-ensemble et chacune des situations biologiques extérieures au bi-ensemble et chacun des gènes extérieurs au bi-ensemble doit être grande.

3.4 Bi-ensembles denses et pertinents

Les contraintes $\mathcal{C}_d$ et $\mathcal{C}_s$ sont complémentaires et peuvent être utilisées conjointement pour augmenter la qualité des motifs extraits.

Etant donné les paramètres α, δ et γ, nous voulons donc calculer les bi-ensembles denses et pertinents, i.e., tous les bi-ensembles satisfaisant $\mathcal{C}_d \wedge \mathcal{C}_s$ dans $\mathbf{r}$. Nous désignons

cette collection par $\mathcal{SAT}_{\alpha\delta\gamma}$. Un bi-ensemble ($S$,$G$) $\in \mathcal{SAT}_{\alpha\delta\gamma}$ ssi :

$$\max_{s\in S}(\sharp\{g \in G \mid (s,g) \notin \mathbf{r}\}) \leq \begin{cases} \alpha \\ |G|/(\gamma+1) \\ \min\limits_{s\in\mathcal{S}\setminus S}(\sharp\{g \in G \mid (s,g) \notin \mathbf{r}\}) - \delta \end{cases}$$

$$\max_{g\in G}(\sharp\{s \in S \mid (s,g) \notin \mathbf{r}\}) \leq \begin{cases} \alpha \\ |S|/(\gamma+1) \\ \min\limits_{g\in\mathcal{G}\setminus G}(\sharp\{s \in S \mid (s,g) \notin \mathbf{r}\}) - \delta \end{cases}$$

Les paramètres α, δ et γ peuvent être différenciés selon que l'on considère ces contraintes sur les lignes et les colonnes. On notera d'un $\prime$ ces paramètres sur les colonnes.

Lorsque $\alpha = \alpha' = 0$, on retrouve des collections déjà bien étudiées :

- $\mathcal{SAT}$ est la collection des 1-rectangles lorsque $\delta = \delta' = 0$.
- $\mathcal{SAT}$ est la collection des itemsets (au sens défini dans la section 2.1) lorsque $\delta = 1$ et $\delta' = 0$.
- $\mathcal{SAT}$ est la collection des objectsets lorsque $\delta = 0$ et $\delta' = 1$.
- $\mathcal{SAT}$ est la collection des concepts formels lorsque $\delta = \delta' = 1$.

Dans le cas où $\alpha = \alpha' = 0$, ces collections correspondent aux bi-ensembles les plus denses et ayant le plus petit seuil de pertinence. Lorsque $\alpha > 0$, les collections de 1-rectangles, d'ensembles et de concepts formels sont généralisées en introduisant un certain nombre d'exceptions (valeur 0) dans les motifs.

La figure 2 montre la collection $\mathcal{SAT}$ lorsque $\alpha = 5$, $\alpha' = 4$, $\delta = \delta' = 1$ et $\gamma = \gamma' = 0$ pour $\mathbf{r}_1$ ordonnée par la relation $\preceq$. Chaque niveau indique le nombre maximum d'exceptions par ligne et par colonne. Par exemple, si une seule exception est autorisée ($\alpha = \alpha' = 1$) et avec $\delta = \delta' = 1$, cinq motifs sont extraits.

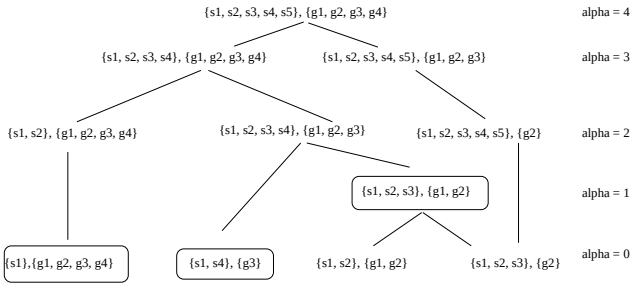

FIG. 2 – Motifs de $\mathbf{r}_1$ avec $\delta = 1$ et $\gamma = 0$. Les motifs entourés sont ceux de $\mathcal{M}_{110}$.

Il peut être pertinent d'étendre les motifs de base (itemset et concepts) avec des exceptions de telle sorte qu'ils conservent les propriétés de maximalité associées à ces motifs au sens de la correspondance de Galois. Cette propriété est très importante car elle permet de mieux appréhender la collection extraite, c'est le cas en particulier pour les biologistes. Pour préserver les correspondances de Galois, nous introduisons une nouvelle contrainte notée $\mathcal{C}_m$.

Définition 6 (Contrainte de maximalité $\mathcal{C}_m$)
Un bi-ensemble (X,Y) $\in \mathcal{SAT}_{\alpha\delta\gamma}$ *satisfait* $\mathcal{C}_m$ *dans* $\mathbf{r}$ *ssi :*
- $\delta = 1$ *et* $\delta' = 0 \Rightarrow \nexists\, (X', Y') \in \mathcal{SAT}_{\alpha\delta\gamma}$ *tel que* $Y = Y'$ *et* $X \subset X'$
- $\delta = 0$ *et* $\delta' = 1 \Rightarrow \nexists\, (X', Y') \in \mathcal{SAT}_{\alpha\delta\gamma}$ *tel que* $X = X'$ *et* $Y \subset Y'$
- $\delta \geq 1$ *et* $\delta' \geq 1 \Rightarrow \nexists\, (X', Y') \in \mathcal{SAT}_{\alpha\delta\gamma}$ *tel que (X,Y)* $\preceq (X', Y')$

La collection des bi-ensembles qui satisfont $\mathcal{C}_d \wedge \mathcal{C}_s \wedge \mathcal{C}_m$ est notée $\mathcal{M}_{\alpha\delta\gamma}$. Sur la figure 2, les trois motifs entourés forment la collection $\mathcal{M}_{110}$. Deux motifs de $\mathcal{SAT}_{110}$ ont été éliminés.

Le tableau 1 montre quelques collections $\mathcal{SAT}$ et $\mathcal{M}$ en fonction des paramètres α et δ.

	$\delta = 1$	
α	$\mathcal{SAT}_{\alpha\delta\gamma}$	$\mathcal{M}_{\alpha\delta\gamma}$
0	$\{\{s_1\},\ \{g_1, g_2, g_3, g_4\}\}$ $\{\{s_1, s_4\}, \{g_3\}\}$ $\{\{s_1, s_2\}, \{g_1, g_2\}\}$ $\{\{s_1, s_2, s_3\}, \{g_2\}\}$	$\{\{s_1\},\ \{g_1, g_2, g_3, g_4\}\}$ $\{\{s_1, s_4\}, \{g_3\}\}$ $\{\{s_1, s_2\}, \{g_1, g_2\}\}$ $\{\{s_1, s_2, s_3\}, \{g_2\}\}$
1	$\{\{s_1\},\ \{g_1, g_2, g_3, g_4\}\}$ $\{\{s_1, s_4\}, \{g_3\}\}$ $\{\{s_1, s_2\}, \{g_1, g_2\}\}$ $\{\{s_1, s_2, s_3\}, \{g_2\}\}$ $\{\{s_1, s_2, s_3\}, \{g_1, g_2\}\}$	$\{\{s_1\},\ \{g_1, g_2, g_3, g_4\}\}$ $\{\{s_1, s_4\}, \{g_3\}\}$ $\{\{s_1, s_2, s_3\}, \{g_1, g_2\}\}$
	$\delta = 2$	
0	$\{\{s_1\},\ \{g_1, g_2, g_3, g_4\}\}$	$\{\{s_1\},\ \{g_1, g_2, g_3, g_4\}\}$
1	$\{\{s_1\},\ \{g_1, g_2, g_3, g_4\}\}$ $\{\{s_1, s_2, s_3\}, \{g_1, g_2\}\}$	$\{\{s_1\},\ \{g_1, g_2, g_3, g_4\}\}$

TAB. 1 – Collections $\mathcal{SAT}_{\alpha\delta\gamma}$ et $\mathcal{M}_{\alpha\delta\gamma}$ sur $\mathbf{r}_1$.

La collection $\mathcal{M}_{\alpha\delta\gamma}$ est muni d'une correspondance de Galois. En effet, dans nos applications, les ensembles de situations permettent d'expliquer l'association des gènes (la co-expression) et inversement. Ainsi, les biologistes recherchent des associations bijectives et décroissantes. Les bi-ensembles extraits vérifient cette propriété.

Propriété 1
Pour $\alpha_1 \leq \alpha$ *et* $\alpha'_1 \leq \alpha'$, δ, δ', γ *et* γ' *fixés, alors* $\forall X \in \mathcal{M}_{\alpha\alpha'\delta\delta'}$, $\exists X_1 \in \mathcal{M}_{\alpha_1\alpha'_1\delta\delta'}$ *tel que* $X_1 \preceq X$. *De plus,* $\forall X_1 \in \mathcal{M}_{\alpha_1\alpha'_1\delta\delta'}$, $\exists X \in \mathcal{M}_{\alpha\alpha'\delta\delta'}$ *tel que* $X_1 \preceq X$.

Propriété 2
Pour α, α', γ *et* γ' *fixés, et* $\delta \leq \delta_1$ *et* $\delta' \leq \delta'_1$ *alors* $\mathcal{SAT}_{\alpha\alpha'\delta_1\delta'_1} \subseteq \mathcal{SAT}_{\alpha\alpha'\delta\delta'}$.

D'après la propriété 1, plus α et α' augmentent, plus la taille de chaque motif extrait de $\mathcal{M}_{\alpha\alpha'\delta_1\delta'_1}$ augmente tout en conservant les associations extraites dans les collections avec α et α' plus petits. En pratique, une réduction importante de la taille de la collection

est observée lorsque les paramètres de l'extraction sont judicieusement choisis (voir section 5). Par conséquent, un effet de zoom est observé lorsque α et α' varient.

Les paramètres δ et δ' permettent de sélectionner les motifs les plus pertinents (voir propriété 2).

$\mathcal{M}_{0,0,0,0}$ et $\mathcal{M}_{0,0,1,1}$ correspondent respectivement aux collections des 1-rectangles et des concepts couvrant toutes les valeurs 1 de la matrice. Ainsi, d'après la propriété 1, $\forall \alpha \geq 0$, $\forall \alpha' \geq 0$ et $\delta, \delta' \in \{0, 1\}$, la collection $\mathcal{M}_{\alpha,\alpha',\delta,\delta'}$ couvre tous les 1 de la matrice.

4 Un algorithme complet

L'algorithme construit un arbre d'énumération binaire, sur les situations biologiques et les gènes, en procédant en profondeur. En s'inspirant du principe de l'algorithme DUAL-MINER (Bucila *et al.*, 2003), chaque nœud de l'arbre est constitué de trois bi-ensembles :

- $O = (O_s, O_g)$ est composé des éléments qui appartiendront aux motifs construits par cette branche,
- $N = (N_s, N_g)$ contient les éléments qui n'appartiendront pas aux motifs engendrés par cette branche,
- $P = (P_g, P_g)$ contient les éléments qui restent à énumérer.

Chaque élément de $\mathcal{S}$ et de $\mathcal{G}$ appartient à un et un seul ensemble parmi O, P et N. Les bi-ensembles O et N sont générés de $(\emptyset, \emptyset)$ au bi-ensemble $(\mathcal{S}, \mathcal{G})$ en exploitant la relation d'ordre $\preceq$.

Pour pouvoir utiliser activement les contraintes $\mathcal{C}_s$ et $\mathcal{C}_d$, on associe à chaque situation biologique s (resp. chaque gène g) deux valeurs notées min_s et max_s (resp. min_g et max_g). min_s correspond au nombre de valeurs 0 de s sur les gènes appartenant à O_g. max_s correspond au nombre de valeurs 0 de s sur les gènes de $O_g \cup P_g$. min_s et max_s correspondent respectivement aux bornes inférieure et supérieure du nombre de 0 à un niveau donné de l'énumération.

4.1 Vérification et propagation des contraintes

A tout moment, les éléments des trois ensembles O, P et N doivent vérifier les contraintes suivantes :

- soit une situation s telle que $\min_s > \alpha$ alors s doit appartenir à N_s. Ainsi, si s était dans O_s, on élague la branche. Sinon s est déplacé dans N_s. En effet, les situations qui ont plus de α valeurs 0 ne peuvent pas appartenir à un bi-ensemble solution.
- soit une situation s telle que

$$max_s < \max_{t \in O_s}\{min_t\} + \delta$$

alors s doit appartenir à O_s. Ainsi, si s appartenait à N_s, le nœud est élagué. Sinon, s est déplacé dans O_s. Dans ce cas, la situation ne contient pas suffisamment de valeurs 0 pour être à l'extérieur du bi-ensemble.

De manière tout à fait similaire, ces contraintes doivent être vérifiées sur les gènes.

D'autres contraintes peuvent également être poussées lors de l'extraction de telle sorte à élaguer l'espace de recherche ou bien à forcer l'appartenance d'un élément à O ou à N. Par exemple, les contraintes monotones et anti-monotones sur $\preceq$ peuvent être exploitées. Les contraintes monotones vont être basée sur $O \cup P$ et les contraintes anti-monotones sur O. Les définitions 7 et 8 donnent des exemples de contraintes.

Définition 7 (Exemple de contraintes monotones)
- $\mathcal{C}_{ms}(\mathbf{r}, \sigma_1, \sigma_2, (S, G))$ *si* $\sharp(O_s \cup P_s) \geq \sigma_1$ *et* $\sharp(O_g \cup P_g) \geq \sigma_2$
- $\mathcal{C}_{Inclusion}(\mathbf{r}, X, Y, (S, G))$ *si* $X \subseteq O_s \cup P_s$ *et* $Y \subseteq O_g \cup P_g$
- $\mathcal{C}_{area}(\mathbf{r}, \sigma, (S, G))$ *si* $\sharp(O_s \cup P_s) * \sharp(O_g \cup P_g) \geq \sigma$

Définition 8 (Exemple de contraintes anti-monotones)
- $\mathcal{C}_{mins}(\mathbf{r}, \sigma_1, \sigma_2, (S, G))$ *si* $\sharp(O_s) \leq \sigma_1$ *et* $\sharp(O_g) \leq \sigma_2$
- $\mathcal{C}_{Inc}(\mathbf{r}, X, Y, (S, G))$ *si* $O_s \subseteq X$ *et* $O_g \subseteq Y$

Si un nœud ne vérifie pas une de ces contraintes alors aucun de ces fils ne la vérifiera et ainsi l'espace de recherche peut être élagué. Ce type d'algorithme permet d'exploiter un grand nombre de contraintes, même des contraintes qui ne sont ni monotones ni anti-monotones sur $\preceq$ comme $\mathcal{C}_d \wedge \mathcal{C}_s$.

4.2 Optimisation

Pour des raisons d'efficacité, nous utilisons une heuristique importante pour l'énumération des gènes et des situations biologiques : l'élément e (gène ou situation biologique) utilisé pour l'énumération est celui qui possède le nombre de valeurs 0 potentiels ($\max_e$) le plus grand. Ce choix tend à réduire la taille du bi-ensemble P le plus rapidement possible. Cela diminue l'espace de recherche tout en préservant la complétude des extractions.

5 Expérimentations

5.1 Evaluation de la robustesse au bruit sur données synthétiques

Pour montrer la pertinence des $\mathcal{M}_{\alpha\alpha'\delta\delta'}$ dans les données bruitées, nous avons tout d'abord généré des jeux de données synthétiques. Notre but est de montrer que l'extraction des $\mathcal{M}_{\alpha\alpha'\delta\delta'}$ permet de retrouver les concepts, introduits dans le jeu de données avant qu'il ne soit bruité. Ainsi, les jeux de données construits sont composés de 4 concepts disjoints comportant chacun 10 éléments sur chaque dimension. Ensuite, un bruit aléatoire uniforme a été introduit dans les données, aussi bien sur les concepts qu'à l'extérieur. Nous avons généré 10 jeux de données pour chaque niveau de bruit : 5%, 10%, 15% et 20%. Le tableau 2 indique le nombre moyen suivi de l'écart-type du nombre de motifs extraits pour chaque niveau de bruit pour $\alpha = \alpha'$ variant de 0 à 3, $\delta = \delta' = 3$ et contenant au moins 4 éléments sur chaque dimension. Ces contraintes permettent de ne pas considérer les petits motifs dus au bruit et de ne conserver que

ceux qui sont très pertinents. Dans le tableau 2, nous donnons également le nombre moyen de concepts pour chaque niveau de bruit.

α		0		1		2		3	
	Nb concepts	Moy	σ	Moy	σ	Moy	σ	Moy	σ
5%	228.6	0	0	1.3	0.82	3.3	0.95	4	0
10%	663.8	0	0	0.1	0.32	1.7	1.16	3	0.94
15%	1292.5	0	0	0	0	0.4	0.70	1.3	0.95
20%	2191.7	0	0	0	0	0	0	3.1	3

TAB. 2 – Moyenne et écart-type du nombre de motifs extraits (sur 10 essais) en fonction de $\alpha = \alpha'$ et du pourcentage de bruit dans les données ($\delta = \delta' = 3$ et $\mathcal{C}_{ms}(\mathbf{r}, 4, 4, (S, G))$).

Lorsqu'il y a 5% de bruit, on retrouve systématiquement les 4 concepts originaux avec $\alpha = \alpha' = 3$. Pour un pourcentage de bruit plus élevé (10% et 15%), seulement certains des concepts originaux sont retrouvés. Lorsque le bruit est trop important (20%), le nombre de motifs extraits est assez variable (l'écart-type vaut 3). Sur certains jeux de données, quelques concepts parmi les 4 d'origine sont retrouvés ; sur d'autres jeux de données, la démultiplication du nombre de concepts réapparaît un peu. En revanche, de très nombreux concepts générés par l'introduction du bruit ont été éliminés.

5.2 Impact des paramètres sur les collections extraites

5.2.1 L'influence des paramètres α et α'

Pour voir l'influence des paramètres α et α' sur $\mathcal{M}_{\alpha\alpha'\delta\delta'}$, nous avons réalisé plusieurs extractions sur le jeu de données CAMDA (Bozdech *et al.*, 2003). Ce jeu de données montre l'évolution des niveaux d'expression de 3719 gènes (colonnes) de Plasmodium falciparum (responsable de la malaria) durant son invasion des globules rouges. La série temporelle comporte 46 mesures du niveau d'expression des gènes.

Nous avons fixé $\delta = \delta' = 1$ et nous avons fait varier $\alpha = \alpha'$ de 0 à 4. De plus, les motifs doivent satisfaire la contrainte $\mathcal{C}_{ms}(\mathbf{r}, \sigma_1, \sigma_2, (S, G))$ avec $\sigma_2 = 3$ et σ_1 qui varie de 19 à 24. Comme la contrainte de fréquence habituellement utilisée lors de l'extraction des ensembles fréquents, la contrainte $\mathcal{C}_{ms}$ permet de rendre les extractions faisables.

Le nombre de motifs extraits pour $\alpha = \alpha'$ de 0 à 2 diminue globalement. Certains motifs sont enrichis et deviennent des sur-ensembles de motifs pour $\alpha = \alpha'$ plus petits. Ensuite, pour $\alpha = \alpha' > 2$, le nombre de motifs extraits tend a augmenter de nouveau. Ceci peut s'expliquer par deux phénomènes :

- Tout d'abord, la taille de certains motifs, initialement non comptabilisés car étant trop petits, augmentent de telle sorte qu'ils satisfont la contrainte de taille
- Lorsque $\alpha \geq 3$, le nombre d'erreurs accepté par ligne est supérieur ou égal au nombre de colonnes minimum du motif, ce qui conduit à accepter des concepts pouvant avoir très peu de 1 par ligne. Cela induit une augmentation du nombre de

α	0	1	2	3	4
$\sigma_1 = 24$	0	4	4	5	5
$\sigma_1 = 23$	9	10	8	9	12
$\sigma_1 = 22$	35	23	22	24	251
$\sigma_1 = 21$	97	68	66	69	-
$\sigma_1 = 20$	241	202	197	213	-
$\sigma_1 = 19$	578	511	513	608	-

TAB. 3 – Nombre de motifs satisfaisant la contrainte $\mathcal{C}_{ms}(\mathbf{r}, \sigma_1, \sigma_2, (S, G))$ avec $\sigma_2 = 3$, σ_1 entre 19 et 24, $\delta = \delta' = 1$ et $\alpha = \alpha'$ qui varie.

motifs. En pratique, il faut imposer une contrainte de taille minimale sur les deux dimensions nettement supérieure à α et α'.

Lorsque α augmente, l'extraction des motifs denses et pertinents devient de plus en plus difficile. Nous n'avons pas réussi à extraire ces motifs pour $\alpha = \alpha' = 4$ et $\sigma_1 \leq 21$.

5.2.2 L'influence des paramètres δ et δ'

Pour montrer l'influence des paramètres δ et δ' sur $\mathcal{M}_{\alpha\alpha'\delta\delta'}$, nous avons réalisé des extractions sur un jeu de données UCI (Internet Advertisements) de dimension 3279×1555. Il ne 'agit pas d'une matrice d'expression mais nous avons cherché un contexte booléen peu dense pour mieux illustrer les variations du nombre de concepts lorsque δ et δ' augmentent.

Pour ces extractions, α et α' sont fixés à 1, δ et δ' varient de 1 à 10 et les motifs extraits (S,G) doivent satisfaire la contrainte $\mathcal{C}_{ms}(\mathbf{r}, \sigma_1, \sigma_2, (S, G))$ avec $\sigma_2 = 0$ et $\sigma_1 \in \{31, 78, 155, 330\}$.

$\delta = \delta'$	1	2	3	4	5	6	7	8	9	10
$\sigma_1 = 31$	549	56	16	7	5	5	2	2	2	2
$\sigma_1 = 78$	131	17	3	2	2	2	1	1	1	1
$\sigma_1 = 155$	43	7	1	1	1	1	1	1	1	1
$\sigma_1 = 330$	6	1	1	1	1	1	1	1	1	1

TAB. 4 – Taille des collections extraites sur le jeu de données de l'UCI :Internet Advertisements, pour $\alpha = \alpha' = 1$ sous la contrainte $\mathcal{C}_{ms}(\mathbf{r}, \sigma_1, \sigma_2, (S, G))$ avec $\sigma_2 = 0$ et $\sigma_1 \in \{31, 78, 155, 330\}$

Les extractions du tableau 4 montrent une diminution importante du nombre de concepts extraits au fur et à mesure de l'augmentation de δ et δ'.

5.3 Extension des concepts

La complexité de l'extraction des motifs denses et pertinents peut augmenter très fortement avec α et α' rendant certaines extractions infaisables. Il est néanmoins possible,

dans ce cas, d'utiliser l'algorithme présenté pour enrichir certains concepts formels intéressant l'utilisateur final. En effet, il suffit pour étendre un concept (S,G) d'extraire les motifs (S', G') de $\mathcal{M}_{\alpha\alpha'\delta\delta'}$ avec α et β supérieur à 0 et tel que (S, G) est un sur-ensemble de (S', G') (il satisfait $\mathcal{C}_{Inclusion}(\mathbf{r}, S', G', (S, G))$). Pour réduire efficacement la complexité du calcul, il faut que le concept que l'on cherche à étendre ait suffisamment d'éléments (relativement à la taille et à la densité du jeu de données utilisé). Dans ce cas, la contrainte d'inclusion devient suffisamment sélective pour réduire l'espace de recherche.

Pour illustrer ce procédé, nous avons utilisé le jeu de données CAMDA qui représente une série temporelle de 46 mesures correspondant à l'évolution du niveau d'expression des 483 gènes dont la fonction biologique est connue parmi 3719 gènes de la matrice d'origine. On peut distinguer trois phases dans le développement de Plasmodium falciparum au cours de l'infection. Elle sont appelées "ring", "trophozoite" et "shizont". Tous les concepts formels ont pu être extraits de cette matrice après discrétisation. Parmi ces 3800 concepts, on s'est intéressé à un concept contenant huit situations relatives à la phase "ring" et quatre gènes dont trois sont connus pour avoir une fonction cytoplasmique. Les gènes ayant cette fonction ont tendance à être sur-exprimés au cours de cette phase. Nous avons essayé d'étendre ce concept pour l'enrichir (voir figure 3). Par exemple en utilisant $\alpha = \alpha' = 2$ et $\delta = \delta' = 1$ on obtient un motif qui contient neuf gènes, onze situations biologiques et 7% de valeurs 0 dans le motif. Les trois situations biologiques ajoutées correspondent à la phase "ring" et parmi les cinq gènes ajoutés, quatre ont une fonction cytoplasmique. Parmi les motifs étendus de la figure 3, cinq des sept nouveaux gènes sont connus pour avoir une fonction cytoplasmique et les huit situations biologiques ajoutées appartiennent à la phase "ring". La prise en compte des exceptions dans les données a permis d'augmenter la taille du motif extrait en ajoutant des éléments cohérents d'un point de vue biologique avec ceux du concept initial.

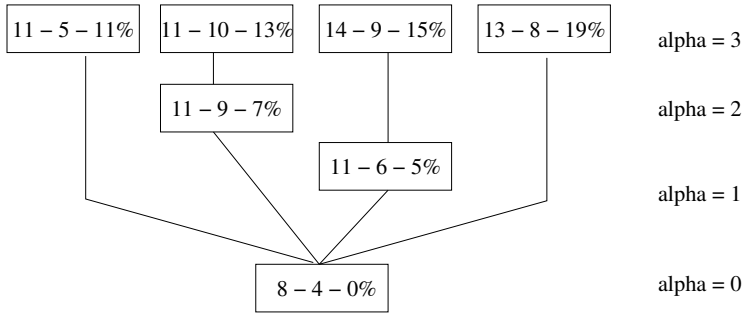

FIG. 3 – Extensions d'un concept : chaque triplet représente le nombre de situations, le nombre de gènes et la densité faible relative de 0.

6 Conclusion

Pour extraire des connaissances dans de grandes matrices booléennes, nous avons défini un nouveau type de motifs appelé bi-ensembles denses et pertinents. Cette recherche a été motivée par des applications en analyse du transcriptome où les concepts formels dans des matrices d'expression de gènes suggèrent aux biologistes des modules de transcription potentiels. Nous nous sommes alors intéressés à la trop grande sensibilité au bruit des extractions de concepts formels pour proposer l'extraction de bi-ensembles qui peuvent être vus comme des concepts formels avec un nombre borné d'exceptions (bi-ensemble dense) mais aussi avec un critère de qualité sur leurs pertinences (singularité des éléments retenus dans le bi-ensemble au regard de l'ensemble des données).

L'extraction de ce nouveau type de motifs est, dans certains cas, plus difficile en pratique que celle de tous les concepts formels. L'applicabilité de l'algorithme complet dans des contextes variés nous paraît donc peu vraisemblable. Pour autant, nous avons proposé une méthode très simple pour exploiter l'algorithme lors de l'extension de certains concepts déjà découverts. Cette direction de recherche nous parait très prometteuse dans l'optique d'une assistance à la découverte de connaissances dans des données réelles, que ce soit dans le cadre de la biologie moléculaire ou plus généralement pour le traitement de données transactionnelles bruitées, denses et/ou très corrélées (i.e., de nombreux domaines d'application où les données sont transactionnelles mais pas le classique contexte de l'analyse du "panier de la ménagère" pour lequel les données sont peu bruitées, peu denses et peu corrélées).

Remerciements

Ce travail est partiellement financé par l'ACI Masse de Données Bingo (MD 46).

Références

AFRATI F. N., GIONIS A. & MANNILA H. (2004). Approximating a collection of frequent sets. In *Proceedings ACM SIGKDD'04*, p. 12–19, Seattle, WA, USA : ACM.

BECQUET C., BLACHON S., JEUDY B., BOULICAUT J.-F. & GANDRILLON O. (2002). Strong association rule mining for large gene expression data analysis : a case study on human SAGE data. *Genome Biology*, **12**. See http ://genomebiology.com/2002/3/12/research/0067.

BESSON J., ROBARDET C. & BOULICAUT J.-F. (2004a). Constraint-based mining of formal concepts in transactional data. In *Proceedings PaKDD'04*, volume 3056 of *LNAI*, p. 615–624, Sydney, Australia : Springer-Verlag.

BESSON J., ROBARDET C. & BOULICAUT J.-F. (2005). *Mining formal concepts with a bounded number of exceptions from transactional data*, In *Post-Workshop proceedings KDID'04*, volume 3377 of *LNCS*, p. 33–45. Springer-Verlag.

BESSON J., ROBARDET C., BOULICAUT J.-F. & ROME S. (2004b). Constraint-based bi-set mining for biologically relevant pattern discovery in microarray data. *Intelligent Data Analysis journal*, **9(1)**. In Press.

BOZDECH Z., LLINÁS M., PULLIAM B. L., WONG E., ZHU J. & DERISI J. (2003). The transcriptome of the intraerythrocytic developmental cycle of plasmodium falciparum. *PLoS Biol*, **1**(e5).

BUCILA C., GEHRKE J. E., KIFER D. & WHITE W. (2003). Dualminer : A dual-pruning algorithm for itemsets with constraints. *Data Mining and Knowledge Discovery*, **7**(4), 241–272.

DHILLON I., MALLELA S. & MODHA D. (2003). Information-theoretic co-clustering. In *Proceedings ACM SIGKDD 2003*, p. 1–10 : ACM.

FU H. & NGUIFO E. M. (2004). Etude et conception d'algorithmes de génération de concepts formels. In *Extraction de motifs dans les bases de données*, volume 9(3/4) of *RSTI série ISI*, p. 109–132. Hermès.

GIONIS A., MANNILA H. & SEPPÄNEN J. K. (2004). Geometric and combinatorial tiles in 0-1 data. In *Proceedings PKDD'04*, volume 3202 of *LNAI*, p. 173–184, Pisa, Italy : Springer-Verlag.

GOETHALS B. & ZAKI M. (2003). *Proceedings of the IEEE ICDM Workshop on Frequent Itemset Mining Implementations FIMI 2003*. Melbourne, USA : IEEE Computer Press.

MEUGNIER E., BESSON J., BOULICAUT J.-F., LEFAI E., DIF N., VIDAL H. & ROME S. (2005). Resolving transcriptional network from microarray data with constraint-based formal concept mining revealed new target genes of SREBP1. *Submitted*.

PASQUIER N., BASTIDE Y., TAOUIL R. & LAKHAL L. (1999). Efficient mining of association rules using closed itemset lattices. *Information Systems*, **24**(1), 25–46.

PEI J., HAN J. & MAO R. (2000). CLOSET an efficient algorithm for mining frequent closed itemsets. In *Proceedings ACM SIGMOD Workshop DMKD'00*.

PENSA R. G., LESCHI C., BESSON J. & BOULICAUT J.-F. (2004). Assessment of discretization techniques for relevant pattern discovery from gene expression data. In *Proceedings ACM BIOKDD'04 co-located with SIGKDD'04*, p. 24–30, Seattle, USA.

RIOULT F., BOULICAUT J.-F., CRÉMILLEUX B. & BESSON J. (2003). Using transposition for pattern discovery from microarray data. In *Proceedings ACM SIGMOD Workshop DMKD'03*, p. 73–79, San Diego, USA.

ROBARDET C. (2002). *Contribution à la classification non supervisée : proposition d'une méthode de bi-partitionnement*. PhD thesis, University Claude Bernard - Lyon 1, F-69622 Villeurbanne cedex.

SEPPÄNEN J. K. & MANNILA H. (2004). Dense itemsets. In *Proceedings ACM SIGKDD'04*, p. 683–688, Seattle, WA, USA : ACM.

STUMME G., TAOUIL R., BASTIDE Y., PASQUIER N. & LAKHAL L. (2002). Computing iceberg concept lattices with titanic. *Data and Knowledge Engineering*, **42**, 189–222.

VENTOS V., SOLDANO H. & LAMADON T. (2004). Treillis de galois alpha. In *Actes CAp 2004*, p. 175–190, Montpellier, F.

WILLE R. (1982). Restructuring lattice theory : an approach based on hierarchies of concepts. In I. RIVAL, Ed., *Ordered sets*, p. 445–470. Reidel.

YANG C., FAYYAD U. & BRADLEY P. S. (2001). Efficient discovery of error-tolerant frequent itemsets in high dimensions. In *Proceedings ACM SIGKDD'01*, p. 194–203, San Francisco, CA, USA : ACM Press.

ZAKI M. J. & HSIAO C.-J. (2002). CHARM : An efficient algorithm for closed itemset mining. In *Proceedings SIAM DM'02*, Arlington, USA.