



**HAL**  
open science

## Extraction de motifs sous contraintes souples de seuil

Willy Ugarte Rojas, Bruno Crémilleux, Samir Loudni, Patrice Boizumault

► **To cite this version:**

Willy Ugarte Rojas, Bruno Crémilleux, Samir Loudni, Patrice Boizumault. Extraction de motifs sous contraintes souples de seuil. JFPC 2012, May 2012, Toulouse, France. hal-00811198

**HAL Id: hal-00811198**

**<https://inria.hal.science/hal-00811198>**

Submitted on 10 Apr 2013

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Extraction de motifs sous contraintes souples de seuil

Willy Ugarte Patrice Boizumault Samir Loudni Bruno Crémilleux

GREYC (CNRS UMR 6072) – Université de Caen Basse-Normandie  
Campus II, Côte de Nacre, 14000 Caen - France  
prenom.nom@unicaen.fr

## Résumé

Lors de l'extraction de motifs sous contraintes, les contraintes sur des mesures (fréquence, taille des motifs) sont parmi les plus utilisées. Dans la pratique, fixer une valeur de seuil pour une mesure est un problème difficile et le choix d'un seuil relève souvent de l'arbitraire. De plus, la rigidité du cadre actuel fait que des motifs pouvant s'avérer en réalité intéressants ne sont pas extraits car ratant de peu le seuil. Dans cet article, nous montrons comment les contraintes souples de seuil peuvent être mises en œuvre dans un extracteur de motifs (basé-CSP). Nous montrons la pertinence et la faisabilité de notre approche au travers d'une application en chémoinformatique portant sur la découverte de fragments moléculaires toxicophores.

## Abstract

Constraint-based pattern discovery is at the core of numerous data mining tasks. Patterns are extracted with respect to a given set of constraints (frequency, closedness, size, etc). In practice, many constraints require threshold values whose choice is often arbitrary. This difficulty is even harder when several thresholds are required and have to be combined. Moreover, patterns barely missing a threshold will not be extracted even if they may be relevant. In this paper, by using CP we propose a method to integrate soft threshold constraints into the pattern discovery process. We show the relevance and the efficiency of our approach through a case study in chemoinformatics for discovering toxicophores.

## 1 Introduction

L'extraction de connaissances dans les bases de données, aussi appelée "fouille de données", a pour but la découverte automatique d'informations nouvelles et interprétables en connaissances utiles. Une étape centrale de ce processus est l'extraction de motifs sous

contraintes, les contraintes permettant de cibler la recherche des informations à extraire suivant les centres d'intérêt de l'utilisateur. Cette forme d'extraction s'est essentiellement développée dans le cas des motifs locaux, ces derniers capturant les phénomènes valides sur une portion de la base de données [8, 12]. Les contraintes sur des mesures telles que la fréquence (nombre d'occurrences d'un motif dans la base de données) ou encore la taille (nombre d'attributs composant un motif) sont parmi les contraintes les plus utilisées.

Dans la pratique, fixer une valeur de seuil pour une mesure est un problème difficile et le choix d'un seuil relève souvent de l'arbitraire. De plus, la rigidité du cadre actuel fait que des motifs pouvant s'avérer en réalité intéressants (i.e. les motifs dont la mesure est proche du seuil mais ne le satisfait pas) ne sont pas examinés. Ce problème est encore plus ardu lorsque plusieurs seuils doivent être fixés simultanément dans une même requête. Une première tentative pour résoudre ce problème, dans le cas des extracteurs de motifs locaux, a été proposée dans [4, 5].

Récemment, plusieurs travaux ont montré l'apport de la Programmation Par Contraintes (PPC) pour l'extraction de motifs locaux [13, 17] ou n-aires [10, 11]. Le point commun de l'ensemble de ces travaux est de modéliser les problèmes d'extraction de motifs, qu'ils soient locaux ou n-aires, sous forme de Problèmes de Satisfaction de Contraintes (CSP). La résolution de ce CSP produit l'ensemble complet des motifs satisfaisant toutes les contraintes.

Dans cet article, nous montrons comment les contraintes souples de seuil peuvent être mises en œuvre dans un extracteur de motifs (basé-CSP) en utilisant les travaux existants en PPC sur la relaxa-

tion de contraintes et les préférences entre solutions. Puis la pertinence et la faisabilité de notre approche sont illustrées au travers d’une application en chimio-informatique portant sur la découverte de fragments moléculaires toxicophores.

Pour cela, à chaque contrainte souple de seuil, est associée une sémantique de violation permettant de mesurer l’écart au seuil. Puis, nous montrons que toute requête incluant des contraintes souples de seuil peut être transformée en une requête équivalente, uniquement constituée de contraintes dures, pouvant être résolues par un solveur de CSP. Nous nous intéressons alors à la recherche des  $k$  meilleurs motifs ( $top_k$ ) selon une mesure d’intérêt [9, 19]. Nous montrons, au travers de l’application à la découverte de fragments moléculaires toxicophores, comment les contraintes souples de seuil permettent d’extraire des motifs (très) pertinents qui n’auraient pas pu être découverts avec des seuils durs.

L’article est organisé comme suit. La section 2 présente le contexte et nos motivations. La section 3 décrit le cadre retenu de la relaxation disjonctive [14, 15] et montre comment les contraintes souples de seuil peuvent être transformées en contraintes dures équivalentes. La section 4 s’intéresse à la recherche des  $top_k$  motifs. La section 5 présente le cadre applicatif de la découverte de fragments moléculaires toxicophores en chimioinformatique. Les expérimentations menées (cf section 6) montrent l’apport des contraintes souples de seuil. La section 7 présente un état de l’art synthétique.

## 2 Contexte et motivations

### 2.1 Définitions

Soit  $\mathcal{I}$  un ensemble de littéraux distincts appelés items. Un motif ensembliste d’items est un sous-ensemble non vide de  $\mathcal{I}$ . Ces motifs sont regroupés dans le langage  $\mathcal{L}_{\mathcal{I}} = 2^{\mathcal{I}} \setminus \emptyset$ . Un contexte transactionnel est un multi-ensemble de motifs de  $\mathcal{L}_{\mathcal{I}}$ . Chaque motif d’un contexte transactionnel constitue une entrée de la base de données et est appelé transaction. La table 1 présente un contexte transactionnel  $\mathcal{T}$  dit “du panier du consommateur” où chaque transaction (panier)  $t_i$  rassemble les articles (items) achetés par le client  $i$ . Les 6 items sont notés  $A, \dots, F$ . Enfin, à chaque article est associé un prix.

L’extraction de motifs sous contraintes a pour objectif d’extraire l’ensemble des motifs de  $\mathcal{L}_{\mathcal{I}}$  présents dans  $\mathcal{T}$  qui satisfont une requête (conjonction de contraintes).

La contrainte de seuil de fréquence permet d’extraire les motifs  $X_i$  dont le nombre d’occurrences dans

Trans.	Items					
$t_1$		$B$			$E$	$F$
$t_2$		$B$	$C$	$D$		
$t_3$	$A$				$E$	$F$
$t_4$	$A$	$B$	$C$	$D$	$E$	
$t_5$		$B$	$C$	$D$	$E$	
$t_6$		$B$	$C$	$D$	$E$	$F$
$t_7$	$A$	$B$	$C$	$D$	$E$	$F$

Items	$A$	$B$	$C$	$D$	$E$	$F$
Prix	30	40	10	40	70	55

TABLE 1 – Exemple de contexte transactionnel  $\mathcal{T}$ .

$\mathcal{T}$  dépasse un seuil minimal fixé par l’utilisateur :  $\text{freq}(X_i) \geq \text{min}_{fr}$ . D’autres mesures quantifient l’intérêt des motifs locaux recherchés. C’est le cas de la taille ou du taux de croissance :

**Définition 1 (taux de croissance)** Soit  $\mathcal{T}$  une base de données partitionnée en deux sous-ensembles  $\mathcal{D}_1$  et  $\mathcal{D}_2$ . Le taux de croissance d’un motif  $X_i$  de  $\mathcal{D}_2$  vers  $\mathcal{D}_1$  est défini par :

$$m_{gr}(X_i) = \frac{|\mathcal{D}_2| \times \text{freq}(X_i, \mathcal{D}_1)}{|\mathcal{D}_1| \times \text{freq}(X_i, \mathcal{D}_2)}$$

**Définition 2 (Jumping Emerging Patterns)**

Soit  $\mathcal{T}$  une base de données partitionnée en deux sous-ensembles  $\mathcal{D}_1$  et  $\mathcal{D}_2$ .  $X_i$  est un Jumping Emerging Pattern (JEP) ssi  $m_{gr}(X_i) = +\infty$  (i.e.  $X_i$  ne figure pas dans  $\mathcal{D}_2$ ).

Par ailleurs, l’utilisateur est très souvent intéressé par la découverte de motifs plus riches que les motifs locaux et qui révèlent des caractéristiques et propriétés de tout l’ensemble des données étudiés. De tels motifs sont appelés motifs n-aires et leur extraction (basée-CSP) est décrite dans [10, 11]. L’approche décrite dans ce papier traite des motifs n-aires.

### 2.2 Motivations à l’aide d’un exemple

**Exemple 2.1** Considérons la requête ci-dessous, permettant d’extraire de la base  $\mathcal{T}$ , tous les motifs fréquents ( $\text{min}_{fr} = 4$ ), de taille au moins 3 et dont la moyenne des prix des articles le composant est supérieure à 45 (mesure  $\text{moyPrix}$ ) :

$$\text{freq}(X_i) \geq 4 \wedge \text{taille}(X_i) \geq 3 \wedge \text{moyPrix}(X_i) \geq 45.$$

Par la suite, nous adoptons la notation  $X_i \langle v_1, v_2, v_3 \rangle$ , où  $X_i$  est un motif, et  $v_1, v_2, v_3$  désignent les valeurs de ces trois mesures pour  $X_i$ .

Sur cet exemple, en considérant uniquement la contrainte de fréquence minimale, il y a 17 solutions.

Avec la conjonction des trois contraintes de la requête, il subsiste une unique solution :  $BDE < 4, 3, 50 >$ .

Considérons les 4 motifs suivants :

- $BEF < 3, 3, 55 >$  -  $BCE < 4, 3, 40 >$
- $CDE < 4, 3, 40 >$  -  $BCDE < 4, 4, 40 >$

Le motif  $BEF$  satisfait deux des trois contraintes de la requête  $Q$ , mais viole légèrement la contrainte de fréquence. Ce motif est clairement intéressant car il présente une dépense moyenne plus élevée que celle du motif  $BDE$ , qui satisfait la requête. Ainsi, en relâchant très légèrement le seuil de fréquence ( $\text{freq}(X_i) \geq 3$ ),  $BEF$  serait solution.

De même, relâcher légèrement le seuil de la contrainte de prix moyen ( $\text{moyPrix}(X_i) \geq 40$ ) permettrait d'extraire trois nouveaux motifs :  $CDE$ ,  $BCE$  et  $BCDE$ . De plus, il est difficile d'affirmer que ces motifs sont nettement moins intéressants que le motif  $BDE$  compte-tenu de l'incertitude quant à la valeur fixée du seuil.

Ainsi, la rigidité du cadre de satisfaction fait qu'un motif potentiellement intéressant n'est pas retenu dès qu'une contrainte de seuil n'est pas vérifiée. Par ailleurs, il n'est pas raisonnable, dans des applications réelles, de considérer que les contraintes sont toutes d'une importance égale. D'où l'idée d'introduire une certaine souplesse, en relaxant les seuils, pour éviter une sélection trop dichotomique des motifs.

### 3 Mise en œuvre des contraintes souples de seuil

Nous présentons la problématique de la relaxation de contraintes ainsi que le cadre général de la relaxation disjonctive [14, 15]. Nous montrons comment les contraintes souples de seuil peuvent être transformées en des contraintes dures équivalentes et être directement résolues par un solveur de CSP.

#### 3.1 Relaxation de contraintes en PPC

Relaxer une contrainte, c'est l'autoriser à ne pas être nécessairement satisfaite contre un coût. Pour cela, on distingue deux catégories de contraintes : les *contraintes d'intégrité* (ou dures) qui doivent être impérativement satisfaites, et les *contraintes de préférence* (ou souples) qui expriment des propriétés que l'on souhaiterait voir vérifiées par une solution. À chaque contrainte souple est associé un coût qui traduit son importance.

Le but de la relaxation n'est plus de satisfaire toutes les contraintes, mais de les satisfaire *au mieux*, i.e. satisfaire toutes les contraintes dures et minimiser la somme des coûts des contraintes souples insatisfaites.

La relaxation se modélise sous forme d'un Problème d'Optimisation sous Contraintes (COP).

Une sémantique de violation  $\mu$  pour une contrainte  $c$  permet de quantifier la violation de  $c$  lorsque  $c$  n'est pas satisfaite. À chaque instanciation  $\mathcal{A}$  des variables de  $c$ , on associe la quantité de violation induite par  $\mathcal{A}$  pour la contrainte  $c$ .

**Définition 3 (sémantique de violation)**  $\mu$  est une sémantique de violation pour la contrainte  $c(X_1, \dots, X_k)$  ssi  $\mu$  est une fonction de  $D_1 \times \dots \times D_k$  vers  $\mathbb{R}^+$  t.q.  $\forall \mathcal{A} \in D_1 \times \dots \times D_k$ ,  $\mu(\mathcal{A}) = 0$  ssi  $c(X_1, \dots, X_k)$  est satisfaite.

Pour une même contrainte, on peut définir plusieurs sémantiques de violation suivant les contextes d'utilisation. Ce sera le cas pour les contraintes de seuil étudiées à la section 3.3, pour lesquelles nous proposons deux sémantiques de violation différentes.

#### 3.2 Cadre général de la relaxation disjonctive

La relaxation disjonctive [14, 15] s'intéresse au problème de satisfaction associé au COP : étant donné une quantité de violation maximale  $\lambda$ , existe-t-il au moins une instanciation de coût inférieur ou égal à  $\lambda$ ? La version relaxée de chaque contrainte est formulée sous forme d'une disjonction : soit la contrainte est satisfaite et le coût est nul, soit la contrainte est insatisfaite et le coût est précisé.

**Définition 4 (relax. disjonctive d'une contrainte)**

Soit  $c$  une contrainte,  $\bar{c}$  sa négation et  $z$  la variable de coût associée. La relaxation disjonctive de  $c$  est la contrainte  $c'$  :

$$c' = [c \wedge (z = 0)] \vee [\bar{c} \wedge (z > 0)]$$

**Exemple 3.1** Soit  $\mathcal{X} = \{X_1, X_2\}$  de domaines  $D_1 = D_2 = \{1, 2, 3\}$ . Soit la contrainte  $X_1 = X_2$  avec comme sémantique de violation  $\mu$  la distance entre  $X_1$  et  $X_2$ , alors  $z = |X_1 - X_2|$ . La relaxation disjonctive de  $c$  est la suivante :

$$c' = [X_1 = X_2 \wedge z = 0] \vee [X_1 \neq X_2 \wedge z = |X_1 - X_2|]$$

Soit  $C_s$  l'ensemble des contraintes souples et  $C_h$  l'ensemble des contraintes dures. À chaque contrainte  $c_i \in C_s$ , on associe une variable de coût  $z_i$ . Soit  $Z$  la variable représentant la violation totale (cumul des violations), alors  $Z = \sum_{c_i \in C_s} z_i$ . Soit  $\lambda$  la quantité maximale de violation que l'on s'autorise. Le problème de satisfaction associé se formule comme suit : Existe-t-il au moins une instanciation telle que  $Z \leq \lambda$ , i.e.  $\sum_{c_i \in C_s} z_i \leq \lambda$ .

Nous avons choisi le modèle de relaxation disjonctive car : (i) le fait que l'on puisse transformer une

contrainte souple en une, ou plusieurs contraintes dures équivalentes, permet de pouvoir traiter la relaxation avec des solveurs de CSP et de bénéficier des avancées faites dans ce domaine; (ii) nous pouvons ainsi directement inclure cette forme de relaxation dans l'extracteur de motifs n-aires (basé CSP) que nous avons précédemment développé [10].

### 3.3 Sémantiques de violation pour les contraintes de seuil

Dans cette section, nous prenons comme exemple introductif la mesure de fréquence, puis nous traitons le cas d'une mesure quelconque.

#### 3.3.1 Mesure de fréquence

Soit  $X_i$  un motif,  $\alpha$  un seuil de fréquence et la contrainte  $c = \mathbf{freq}(X_i) \geq \alpha$ . Une première sémantique de violation  $\mu_1$  consiste à mesurer l'écart absolu au seuil. Mais, pour pouvoir cumuler les violations des différentes contraintes de seuil, il est nécessaire de travailler avec des écarts relatifs. Une seconde sémantique de violation  $\mu_2$  consiste à associer, à chaque motif  $X_i$ , l'écart relatif de sa fréquence au seuil  $\alpha$  :

$$\mu_2(X_i) = \begin{cases} 0 & \text{si } \mathbf{freq}(X_i) \geq \alpha \\ \frac{\alpha - \mathbf{freq}(X_i)}{\alpha} & \text{sinon} \end{cases}$$

#### 3.3.2 Mesure $m$ quelconque

Soit  $\mathcal{I}$  un ensemble d'items et  $\mathcal{T}$  un ensemble de transactions. Soit  $max_m$  la valeur maximale<sup>1</sup> pour la mesure  $m$ .

si  $c = m(X_i) \geq \alpha$  alors

$$\mu_2(X_i) = \begin{cases} 0 & \text{si } m(X_i) \geq \alpha \\ \frac{\alpha - m(X_i)}{\alpha} & \text{sinon} \end{cases}$$

si  $c = m(X_i) \leq \alpha$  alors

$$\mu_2(X_i) = \begin{cases} 0 & \text{si } m(X_i) \leq \alpha \\ \frac{m(X_i) - \alpha}{max_m - \alpha} & \text{sinon} \end{cases}$$

### 3.4 Transformation des contraintes souples de seuil

Cette section montre comment transformer des contraintes souples de seuil en contraintes dures équivalentes. Tout d'abord, nous modélisons le problème d'extraction des motifs n-aires sous forme de CSP.

1. Pour la fréquence,  $max_m = |\mathcal{T}|$ ; pour la taille,  $max_m = |\mathcal{I}|$ .

Puis, nous présentons la transformation en deux temps : un exemple introductif (fréquence) puis une mesure quelconque. Enfin, nous décrivons le CSP résultant associé à la relaxation dans le cadre disjonctif.

#### 3.4.1 CSP initial

Soit  $\mathcal{I}$  l'ensemble des items,  $\mathcal{T}$  l'ensemble des transactions. Tout problème d'extraction de motifs n-aires peut se modéliser [10, 11] sous la forme d'un CSP  $\mathcal{P} = (\mathcal{X}, \mathcal{D}, \mathcal{C})$  tel que :

- $\mathcal{X} = \{X_1, \dots, X_n\}$ , chaque variable  $X_i$  représente un motif inconnu.
- $\mathcal{D} = \{D_{X_1}, \dots, D_{X_n}\}$ , chaque domaine  $D_{X_i}$  est l'intervalle ensembliste  $[\emptyset .. \mathcal{I}]$ .
- $\mathcal{C}$  est l'ensemble des contraintes que l'on peut partitionner en  $\mathcal{C}_{ens} \cup \mathcal{C}_{num}$  où :
  - $\mathcal{C}_{ens}$  est un ensemble de contraintes ensemblistes. Exemples :  $X_1 \subset X_2, I \in X_4, \dots$
  - $\mathcal{C}_{num}$  est un ensemble de contraintes numériques sur les mesures. Exemples :  $|\mathbf{freq}(X_1) - \mathbf{freq}(X_2)| \leq \alpha_1, \mathbf{size}(X_4) < \mathbf{size}(X_1) + 1, \dots$

#### 3.4.2 Mesure de fréquence

Soit  $X_i$  un motif,  $\alpha$  un seuil de fréquence et la contrainte  $c = \mathbf{freq}(X_i) \geq \alpha$ . Soit  $z$  la variable de coût associée. La relaxation disjonctive de  $c$  pour  $\mu_2$  est :

$$[(\mathbf{freq}(X_i) \geq \alpha) \wedge z = 0] \vee [(\mathbf{freq}(X_i) < \alpha) \wedge z = \frac{\alpha - \mathbf{freq}(X_i)}{\alpha}]$$

Que l'on peut reformuler de manière équivalente par l'unique contrainte :  $z = \max(0, \frac{\alpha - \mathbf{freq}(X_i)}{\alpha})$

#### 3.4.3 Mesure $m$ quelconque

En appliquant une transformation identique à celle de la mesure de fréquence, on obtient la reformulation, en contraintes dures équivalentes, des contraintes de seuil souples associées à  $m$ .

- La relaxation de  $c = (m(X_i) \geq \alpha)$  est :  
 $c' = [z = \max(0, \frac{\alpha - m(X_i)}{\alpha})]$
- La relaxation de  $c = (m(X_i) \leq \alpha)$  est :  
 $c' = [z = \max(0, \frac{m(X_i) - \alpha}{max_m - \alpha})]$

Ainsi, toute requête contenant une ou plusieurs contraintes souples de seuil  $c_i$  peut être transformée en une requête *équivalente* ne contenant que des contraintes dures : si  $c_i$  est une contrainte dure alors elle reste telle quelle ; si  $c_i$  est une contrainte souple de seuil alors elle est remplacée par sa transformée.

Enfin, soit  $\lambda$  la quantité maximale de violation autorisée. On définit la variable de coût  $Z = \sum_{c_i} z_i$  représentant le cumul des violations, où  $z_i$  est la variable

de coût associée à chaque contrainte souple de seuil  $c_i$ . Enfin, on ajoute la contrainte  $Z \leq \lambda$ .

### 3.4.4 CSP issu de la transformation

Soit  $\lambda$  la quantité maximale de violation autorisée ( $\lambda$  compris entre 0 et 100%). Soit  $\mathcal{P}' = (\mathcal{X}', \mathcal{D}', \mathcal{C}')$  le CSP obtenu par la relaxation disjonctive du CSP initial  $\mathcal{P} = (\mathcal{X}, \mathcal{D}, \mathcal{C})$  :

- $\mathcal{X}' = \mathcal{X} \cup_{1 \leq i \leq k} \{z_i\} \cup \{Z\}$ ,
- $\mathcal{D}' = \mathcal{D} \cup_{1 \leq i \leq k} \{D_{z_i}\} \cup \{D_Z\}$  avec  $D_{z_i} = [0..100]$  et  $D_Z = [0..\lambda]$  qui sont des intervalles d'entiers,
- $\mathcal{C}' = \mathcal{C}_{ens} \cup \mathcal{C}'_{num} \cup \{Z = \sum_{1 \leq i \leq k} z_i\}$  avec  $\mathcal{C}'_{num} = \mathcal{C}_{hard} \cup \mathcal{C}_{disj}$  où :
  - $\mathcal{C}_{hard}$  est l'ensemble des contraintes numériques dures,
  - $\mathcal{C}_{disj}$  est l'ensemble des contraintes dures associées aux  $k$  contraintes souples de seuil.

## 4 Extraction des k-meilleurs motifs

En fouille de données, la recherche des  $k$  meilleurs motifs selon une mesure d'intérêt ( $top_k$  motifs) se révèle très utile pour trouver les motifs les plus significatifs au regard d'un critère choisi par l'utilisateur. Plusieurs méthodes adaptées à différentes mesures d'intérêt ont été proposées dans cette direction [9, 19].

Pour cela, on définit une mesure d'intérêt sur les motifs extraits, i.e. ceux satisfaisant la requête initiale, de façon à pouvoir les comparer.

### 4.1 Exemple introductif

Soit  $X_i$  un motif,  $\mu$  une sémantique de violation et la contrainte de seuil  $\text{freq}(X_i) \geq \alpha$ . Une sémantique de violation ne permet pas de prendre en compte le degré de satisfaction d'une contrainte  $c$ . En effet, si un motif  $X_i$  satisfait une contrainte  $c$  alors  $\mu(X_i) = 0$ . Or un motif  $X_i$ , dont la fréquence est très grande par rapport au seuil  $\alpha$ , sera considéré comme plus intéressant qu'un motif  $X_j$  dont la fréquence est légèrement supérieure à  $\alpha$ .

### 4.2 Intérêt d'un motif pour une contrainte de seuil

Soit  $m$  une mesure, et  $max_m$  la valeur maximale pour cette mesure. Nous proposons la mesure d'intérêt  $\theta_m : \{X_i \in \mathcal{L}_{\mathcal{I}}, \mu_2(X_i) \leq \lambda\} \rightarrow [-100..100]$  définie par :

$$si \quad c = m(X_i) \geq \alpha \quad alors$$

$$\theta_m(X_i) = \begin{cases} \frac{m(X_i) - \alpha}{max_m - \alpha} & si \quad m(X_i) \geq \alpha \\ \frac{m(X_i) - \alpha}{\alpha} & sinon \end{cases}$$

si  $c = m(X_i) \leq \alpha$  alors

$$\theta_m(X_i) = \begin{cases} \frac{\alpha - m(X_i)}{\alpha} & si \quad m(X_i) \leq \alpha \\ \frac{\alpha - m(X_i)}{max_m - \alpha} & sinon \end{cases}$$

### 4.3 Intérêt d'un motif pour une requête

Considérons à présent un ensemble de mesures  $\mathcal{M}$  et une requête exprimée sous la forme d'une conjonction de contraintes de seuil de la forme  $m(X_i) \geq \alpha_m$  (ou bien  $\leq$ ). On définit l'intérêt d'un motif pour une requête (conjonction de contraintes) comme étant le cumul des intérêts des contraintes qui la composent :

$$\theta(X_i) = \sum_{m \in \mathcal{M}} \gamma_m \times \theta_m(X_i)$$

où  $\gamma_m$  est un coefficient traduisant l'importance de la mesure  $m$ . On peut alors extraire les  $top_k$  motifs, i.e. les  $k$  meilleurs motifs selon la mesure d'intérêt  $\theta$ .

## 5 Découverte de fragments toxicophores

La toxicologie est la science étudiant les substances chimiques toxiques. Elle s'intéresse notamment à l'identification de fragments moléculaires spécifiques appelés toxicophores et considérés comme responsables des propriétés toxiques d'une substance chimique. Un objectif majeur est alors la découverte de tels fragments afin de mieux identifier les caractéristiques des molécules liées à la toxicité.

### 5.1 Fragments toxicophores

Un motif chimique émergent est une conjonction de fragments moléculaires qui apparaît fréquemment dans une classe de molécules et peu fréquemment dans une autre classe [1, 2]. L'émergence d'un motif chimique est mesurée par le taux de croissance entre les deux classes (cf définition 1). La combinaison des mesures d'émergence et de fréquence (afin d'assurer une certaine représentativité), s'avère précieuse pour la prédiction de la toxicité [16]. Par ailleurs, d'autres mesures issues des connaissances chimiques, comme l'aromaticité ou la densité d'une molécule, sont aussi des indicateurs connus de la toxicité. C'est pourquoi nous avons exploité ces différents types de mesures pour l'extraction des toxicophores (cf section 5.2).

Nous avons utilisé un jeu de données (base European Chemicals Bureau) préparé par le CERMN<sup>2</sup>. La toxicité est fondée sur l'indicateur quantitatif de toxicité

2. Centre d'Etudes et de Recherche sur le Médicament de Normandie, UPRES EA 4258 FR CNRS 3038, Université de Caen Basse-Normandie.

$CL50^3$ . En fonction de la valeur de cet indicateur, les molécules sont réparties dans les trois catégories suivantes :  $H400$  très toxique ( $CL50 \leq 1$  mg/L),  $H401$  toxique ( $1$  mg/L  $< CL50 \leq 10$  mg/L), et  $H402$  nocif ( $10$  mg/L  $< CL50 \leq 100$  mg/L).

Dans cette étude, nous nous concentrons uniquement sur les classes  $H400$  et  $H402$ . Le jeu de données utilisé contient 567 molécules, 372 de la classe  $H400$  ( $\mathcal{D}_1$ ) et 195 de la classe  $H402$  ( $\mathcal{D}_2$ ). Les molécules sont représentées en utilisant 129 sous-graphes connexes fréquents initialement extraits au seuil de fréquence de 10% [16].

## 5.2 Contraintes de seuil considérées

**L'émergence** permet de caractériser une molécule d'une classe (toxique) par rapport à une autre classe (non-toxique). Les motifs émergents traduisent l'hypothèse toxicophore (**H1**) : si une molécule possède dans sa structure les fragments moléculaires d'un motif émergent, alors elle possède des caractéristiques de toxicité et est donc particulièrement susceptible d'être toxique. L'émergence est mesurée par le taux de croissance  $m_{gr}$  (cf définition 1). Soit  $min_{gr}$  un seuil minimal pour le taux de croissance. On impose la contrainte souple de seuil :  $m_{gr}(X_i) \geq min_{gr}$ .

**Fréquence.** Les motifs de faible fréquence sont souvent dus à des artefacts dans les données et constituent du bruit. Afin d'assurer une représentativité de l'information extraite, on impose la contrainte souple de seuil :  $freq(X_i) \geq min_{fr}$ , où  $min_{fr}$  est un seuil minimal pour la fréquence.

**Aromaticité.** L'intérêt de cette mesure est qu'elle véhicule une hypothèse toxicophore (**H2**) : plus la valeur d'aromaticité est forte, plus la molécule possédant ces fragments moléculaires tend à être toxique. L'aromaticité d'un motif est la moyenne de l'aromaticité de ses fragments moléculaires. Soit  $m_a$  la mesure d'aromaticité d'un motif. On impose la contrainte souple de seuil :  $m_a(X_i) \geq min_a$ , où  $min_a$  est un seuil minimal pour l'aromaticité.

**Densité.** Plus un sous-graphe codant une molécule est dense<sup>4</sup>, plus son comportement chimique est fort. Un motif composé de fragments moléculaires denses renforce l'hypothèse toxicophore (**H3**). La densité d'un motif est la moyenne des densités de ses fragments.

3. Concentration nécessaire d'une substance pour causer la mort de 50% d'une population dans des conditions expérimentales précises.

4. La densité d'un sous-graphe est égale à  $2e/v(v-1)$ , où  $e$  est son nombre d'arêtes et  $v$  son nombre de sommets.

Soit  $m_d$  la mesure de densité et  $min_d$  un seuil minimal. On impose la contrainte souple de seuil :  $m_d(X_i) \geq min_d$

## 6 Expérimentations

### 6.1 Protocole expérimental

La requête soumise  $q(X_i)$  est la conjonction des quatre contraintes souples de seuil présentées à la section 5.2.

Pour l'aromaticité et la densité, les seuils ont été fixés à environ 2/3 de leur valeur maximale ( $min_a=60$  et  $min_d=60$ ). En effet, des seuils élevés sont en faveur des hypothèses toxicophores (**H2**) et (**H3**). Pour l'émergence et la fréquence, les seuils ont été fixés à environ 1/4 de leur valeur maximale ( $min_{gr}=5$  et  $min_{fr}=90$ ) de manière à obtenir un compromis entre la fréquence et le taux de croissance. Un tel choix nous permet de n'extraire que les motifs les plus fréquents ayant les meilleurs taux de croissance.

Nous avons retenu la sémantique de violation  $\mu_2$  (écart relatif) car les contraintes sont de nature hétérogène. Nous avons considéré trois valeurs différentes pour la quantité maximale de violation autorisée :  $\lambda \in \{0, 20\%, 40\%\}$ . Pour évaluer l'intérêt des motifs extraits, nous avons fixé  $\gamma_{gr}$ ,  $\gamma_{fr}$  et  $\gamma_d$  à 1 et  $\gamma_a$  à 2. En effet, l'aromaticité constitue une connaissance chimique plus importante.

Pour évaluer la présence de toxicophores dans les motifs émergents de fragments moléculaires extraits, nous avons répertorié six fragments moléculaires ayant des propriétés écotoxicologiques connues, comme le benzène, le phénol, le chlorobenzène, les organophosphorés, l'aniline et le pyrrole.

Toutes nos expérimentations ont été réalisées sur un processeur Intel core i3 à 2,13 GHz ayant 4 Go de RAM. La mise en œuvre de notre approche a été réalisée en **Gecode** par extension de l'extracteur de motifs n-aires basé-CSP développé par M. Khiari [10].

### 6.2 Extraction des motifs émergents

Le tableau 2 indique les nombres de motifs émergents extraits contenant au moins un toxicophore dans son intégralité (colonnes notées **T**) ou des sous-fragments d'un toxicophore (colonnes notées **F**) parmi cinq des six fragments moléculaires répertoriés dans la base, ceci pour les trois valeurs de  $\lambda$  retenues (cf section 6.1).

Cette répartition est donnée pour le nombre total de solutions trouvées (col. 2-7) mais aussi pour les  $top_{25}$  (col 8-13) et les  $top_{50}$  (col 14-19). Comme les deux catégories **T** et **F** ne sont pas disjointes, la somme de leurs nombres de motifs est toujours supérieure ou

$\lambda$	Total						top - 25						top - 50					
	0		20		40		0		20		40		0		20		40	
	7650		402204		4289335		28		37		57		55		64		85	
# Solutions	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F
Benzène c1ccccc1 	1912	7573	183881	396749	1565883	4210482	0	25	<b>2</b>	25	<b>6</b>	24	7	50	<b>7</b>	50	<b>8</b>	49
Phénol c1(ccccc1)O 	900	4519	93632	217195	556890	3234279	2	9	<b>6</b>	3	<b>2</b>	0	4	18	<b>9</b>	12	<b>5</b>	8
Chlorobenzène Clc1ccccc1 	0	3041	74182	184502	253429	509281	0	14	<b>2</b>	14	<b>2</b>	1	0	28	<b>7</b>	22	<b>2</b>	15
Aniline c1(ccccc1)N 																		
Pyrrole c1cnc1 						1						1						1

TABLE 2 – Répartition des motifs émergents en fonction des toxicophores connus.

égale au nombre total de solutions (# solutions). Le temps nécessaire pour extraire l'ensemble de toutes les solutions est de 21 s. pour ( $\lambda=0$ ), 9 min. pour ( $\lambda=20$ ) et 6h15 min. pour ( $\lambda=40$ ).

Comme le montrent les résultats du tableau 2, 45%<sup>5</sup> (resp. 36.5%) des solutions extraites avec  $\lambda=20$  (resp. 40) contiennent du benzène (fragment de type **T**), contre environ 25% pour  $\lambda=0$ . Les seuils souples permettent ainsi de mieux retrouver ce toxicophore (gain moyen d'environ 16%). Pour les fragments de type **F**, la proportion de solutions extraites contenant des sous-fragments du benzène ( $\{cc, ccc, cccc, ccccc\}$ ) est quasi identique dans le cas dur et le cas souple (environ 98%). Cette tendance se confirme sur le phénol, où 23% (resp. 13%) des solutions extraites avec  $\lambda=20$  (resp. 40%) contiennent un tel fragment, contre 11% pour  $\lambda=0$ . De nouveau, les seuils souples permettent de mieux retrouver ce toxicophore (gain moyen d'environ 7%).

Pour le chlorobenzène (avec  $\lambda=0$ ), seuls les motifs contenant des fragments moléculaires de type **F** sont extraits :  $\{Clc(cc)cc, Clc(cc)ccc, Clc(cc)cccc, Clc(cc)ccc, Clccc\dots\}$ . Les seuils souples permettent de retrouver en moyenne 19% de toxicophores contenant du chlorobenzène (i.e, fragment de type **T**). De plus, pour le pyrrole, les seuils souples permettent de détecter un nouveau motif contenant le sous-fragment *nc*. Sans les seuils souples, ce motif serait difficile à extraire car associé à un sous-fragment moléculaire ayant une valeur de fréquence relativement faible.

Les motifs contenant de l'aniline ne sont pas détectés en raison de leur faible densité (33). En effet, pour  $\lambda=40$ , la valeur minimale autorisée est de

$60 \times 0.60 = 36$ . Une augmentation très légère ( $\lambda=45$ ), permettrait d'extraire ces motifs. Enfin, les motifs organo-phosphorés sont caractérisés par une très forte émergence ( $+\infty$ ). Ce sont des JEPs (cf définition 2) : ils ne figurent pas dans le tableau 2 et sont traités à part dans la section 6.4).

### 6.3 Extraction des $top_k$ motifs

Les résultats du tableau 2 montrent que sur les  $top_{25}$  (resp.  $top_{50}$ ) motifs extraits avec  $\lambda=0$ , seuls 2 (resp. 4) motifs contiennent du phénol. Par ailleurs, les  $top_k$  motifs extraits sont constitués uniquement de sous-fragments de benzène ou de chlorobenzène.

Le tableau 3 donne les  $top_{25}$  motifs émergents extraits avec  $\lambda=20$ . Les lignes jaunes indiquent les motifs extraits avec  $\lambda=0$  et contenant du phénol dans son intégralité, alors que les lignes grisées correspondent aux nouveaux motifs obtenus grâce aux contraintes souples de seuil (les seuils violés sont sur-lignés en noir).

Les seuils souples permettent de retrouver 4 nouveaux motifs contenant du phénol sur les  $top_{25}$  extraits (lignes 17 – 20), ce qui représente un ratio de 3 ( $\lambda=20$  permet de détecter 3 fois plus de meilleures solutions comparé à  $\lambda=0$ ). Notons que 2 de ces motifs contiennent également du benzène (lignes 18 et 20). Par ailleurs, ces motifs, qui violent très légèrement la contrainte de densité, sont caractérisés par une très forte aromaticité et une forte émergence, ce qui renforce nos hypothèses toxicophores sur l'émergence (**H1**) et l'aromaticité (**H2**). De plus,  $\lambda=20$  permet d'extraire 2 nouveaux motifs contenant du chlorobenzène (lignes 2 et 4) et un motif contenant le fragment  $Clc(cc)ccc$  (ligne 10). Ces motifs sont aussi intéressants, car ils renforcent les hypothèses toxicophores précédemment émises.

5. Ratio entre le nombre de solutions contenant un toxicophore et le nombre total de solutions.

N	Intérêt	Motif				Émergence	Fréquence	Aromaticité	Densité	SMILES <sup>6</sup>				Condensée		
		1	2	3	4					1	2	3	4			
1	193	24	35	69	7	101	95	66	cc	ecc	c1(ccccc1)O		c1(ccccc1)O			
2	191	13	24	35	8	89	95	66	Clc1ccccc1	cc	ccc		Clc1ccccc1			
3	189	24	35	47	69	7	101	96	62	cc	ccc	cccc	c1(ccccc1)O	c1(ccccc1)O		
4	187	13	24	35	47	8	89	96	62	Clc1ccccc1	cc	ccc	cccc	Clc1ccccc1		
5	185	12	24	35	47	8	90	96	61	Clc(c)cccc	cc	ccc	cccc	Clc(c)cccc		
6	185	14	24	35	47	8	90	96	61	Clc1ccccc1	cc	ccc	cccc	Clc1ccccc1		
7	185	24	35	47	68	6	103	96	61	cc	ccc	cccc	cccc(c)O	cccc(c)O		
8	185	24	35	47	80	6	103	96	61	cc	ccc	cccc	ccc(cc)O	ccc(cc)O		
9	185	24	35	38		5	118	90	72	cc	ccc	cccO		cccO		
10	184	24	35	47	78	8	89	96	61	cc	ccc	cccc	Clc(cc)ccc	Clc(cc)ccc		
11	184	6	24	35		9	93	90	72	Clc(c)c	cc	ccc		Clc(c)c		
12	184	8	24	35	47	9	93	94	64	Clc(c)cc	cc	ccc	cccc	Clc(c)cc		
13	184	8	24	35		9	93	92	68	Clc(c)cc	cc	ccc		Clc(c)cc		
14	184	7	24	35		9	94	90	72	Clccc	cc	ccc		Clccc		
15	184	9	24	35	47	9	94	94	64	Clccc	cc	ccc	cccc	Clccc		
16	184	9	24	35		9	94	92	68	Clccc	cc	ccc		Clccc		
17	183	24	35	47	59	69	7	101	97	57	cc	ccc	cccc	cccc	c1(ccccc1)O	c1(ccccc1)O
18	183	24	35	47	69	77	7	101	97	57	cc	ccc	cccc	c1(ccccc1)O	c1ccccc1	c1(ccccc1)O
19	183	24	35	59	69		7	101	96	59	cc	ccc	cccc	c1(ccccc1)O		c1(ccccc1)O
20	183	24	35	69	77		7	101	96	59	cc	ccc	c1(ccccc1)O	c1ccccc1		c1(ccccc1)O
21	183	12	24	35		8	90	94	64	Clc(c)cccc	cc	ccc		Clc(c)cccc		
22	183	14	24	35		8	90	94	64	Clc1ccccc1	cc	ccc		Clc1ccccc1		
23	183	11	24	35	47		9	92	95	62	Clc1ccccc1	cc	ccc	cccc	Clc1ccccc1	
24	183	11	24	35		9	92	93	66	Clc1ccccc1	cc	ccc		Clc1ccccc1		
25	183	10	24	35	47		9	93	95	62	Clc(c)ccc	cc	ccc	cccc	Clc(c)ccc	

<sup>6</sup><http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html>

TABLE 3 –  $top_{25}$  motifs émergents extraits avec  $\lambda=20$ .

Le tableau 4 montre les  $top_{25}$  motifs émergents extraits avec  $\lambda=40$ . Comme précédemment, les contraintes souples de seuil permettent de retrouver 6 nouveaux motifs contenant du benzène (cf. lignes 7, 9, 14, 16, 19 et 20). Ces motifs, qui violent très légèrement le seuil d'émergence, sont très fortement aromatiques et relativement denses, ce qui renforce de nouveau l'hypothèse toxicophore liée à la densité (H3). Un nouveau motif particulièrement intéressant pour les chimistes est obtenu :  $\{nc\}$ . Ce motif, incluant un sous-fragment du pyrrole, est réputé dangereux pour l'environnement car il est très toxique pour les espèces aquatiques.

Par ailleurs, si l'on considère les  $top_{50}$  motifs extraits, les seuils souples  $\lambda=20$  (resp. 40) permettent de détecter 2.25 (resp. 1.25) fois plus de meilleures solutions contenant du phénol. De plus,  $\lambda=40$  permet d'extraire 8 (resp. 2) nouveaux motifs contenant du benzène (resp. chlorobenzène). Ainsi, l'ensemble de ces résultats confirme l'intérêt des seuils souples pour l'extraction de motifs toxicophores.

## 6.4 Extraction des Jumping Emerging Patterns

Cette seconde série d'expérimentations évalue le caractère de toxicité porté par les fragments moléculaires présents uniquement dans les molécules très toxiques (classe H400), autrement dit les Jumping Emerging Patterns (cf définition 2). Le tableau 5, présente les  $top_{25}$  JEPs extraits pour différentes valeurs de  $\lambda$ . On peut dresser les remarques suivantes :

(i) Sans les seuils souples, aucun JEP n'a pu être extrait.

(ii) Avec  $\lambda=50$  (resp. 60), nous obtenons 3 (resp. 457) JEPs. En effet, ces motifs sont peu fréquents, il est donc nécessaire d'avoir un seuil de violation relativement élevé.

(iii) Tous les motifs contenant des sous-fragments organo-phosphorés ont un taux de croissance infini. Ce composant est une généralisation de plusieurs Jumping Emerging Fragments et peut être vu comme une structure commune maximale de ces fragments.

(iv) Parmi les  $top_{25}$  motifs extraits avec  $\lambda=60$ , les motifs les plus intéressants sont ceux contenant un *cycle benzénique* ( $c1ccccc1$ ). En effet, le benzène est un fragment moléculaire très aromatique. Pour  $\lambda=50$ , seuls les motifs contenant des sous-fragments du benzène sont extraits. Ces motifs sont malgré tout moins pertinents d'un point de vue chimique.

A nouveau, ces résultats montrent la pertinence et l'apport des contraintes souples de seuil pour mettre en évidence les structures chimiques les plus parlantes, comme par exemple les cycles benzéniques par rapport à leurs sous-fragments.

## 7 Travaux relatifs

En fouille de données, la relaxation a été utilisée pour obtenir des contraintes relaxées ayant des propriétés de monotonie dans le but de réutiliser les algorithmes usuels de filtrage. Ainsi, les contraintes basées sur des expressions régulières sont relaxées en des contraintes anti-monotones pour extraire des séquences [7]. Dans [18], les auteurs proposent de générer automatiquement une relaxation monotone ou anti-monotone d'une contrainte.

N	intérêt	motif				émergence	fréquence	aromaticité	densité	SMILES	Condensée
1	301	24				3	289	100	100	cc	cc
2	275	15				7	65	100	100	nc	nc
3	258	24	35			3	288	100	83	cc	ccc
4	237	24	47			3	281	100	75	cc	cccc
5	230	24	35	47		3	281	100	72	cc	ccc
6	224	24	59			3	279	100	70	cc	cccc
7	223	24	77			3	274	100	70	cc	c1cccc1
8	219	24	35	59		3	279	100	68	cc	ccc
9	218	24	35	77		3	274	100	68	cc	ccc
10	216	35				3	288	100	65	ccc	ccc
11	213	24	35	76		3	274	100	66	cc	ccc
12	213	24	76			3	274	100	66	cc	cccc
13	209	24	35	47	59	3	279	100	64	cc	ccc
14	208	24	35	47	77	3	274	100	64	cc	ccc
15	206	24	47	59		3	279	100	63	cc	ccc
16	205	24	47	77		3	274	100	63	cc	ccc
17	203	24	35	47	76	3	274	100	62	cc	ccc
18	200	24	47	76		3	274	100	61	cc	ccc
19	200	24	35	59	77	3	274	100	61	cc	ccc
20	198	24	59	77		3	274	100	60	cc	ccc
21	193	24	35	69		7	101	95	66	cc	ccc
22	191	13	24	35		8	89	95	66	C1c1cccc1	cc
23	189	24	35	47	69	7	101	96	62	cc	ccc
24	187	13	24	35	47	8	89	96	62	C1c1cccc1	cc
25	185	12	24	35	47	8	90	96	61	C1c(c)cccc	cc

TABLE 4 –  $top_{25}$  motifs émergents extraits avec  $\lambda=40$ .

Pour les motifs locaux, [4, 5] ont proposé un cadre formel pour l’expression de préférences entre solutions. Chaque contrainte possède sa propre mesure d’intérêt et l’intérêt d’une requête est une agrégation des intérêts des contraintes qui la composent. Étant donnée une requête, il s’agit de trouver tous les motifs (locaux) dont la mesure d’intérêt satisfait un seuil.

Mais, cette approche repose sur une hypothèse très forte : l’intérêt d’une requête satisfait un seuil, si et seulement si, l’intérêt de *chaque* contrainte satisfait ce même seuil [4, 5]. Si on agrège les coûts par l’opérateur *min* (cas des *fuzzy sets*), alors on a bien l’équivalence. Par contre, dans le cadre additif que nous traitons (et aussi dans le cadre probabiliste), ce n’est plus le cas. C’est pourquoi les auteurs ont besoin de faire d’une étape supplémentaire (post-traitement) pour ne garder que les solutions satisfaisant le seuil.

Ainsi, à la différence de [4, 5], l’approche que nous proposons (i) préserve l’équivalence et ne nécessite donc aucun post-traitement, et (ii) elle s’applique aux motifs n-aires, et donc aux motifs locaux (unaires).

## 8 Conclusion

Nous avons proposé un cadre général permettant de mettre en œuvre les contraintes souples de seuil dans un extracteur de motifs (basé-CSP) en utilisant les travaux en PPC sur la relaxation de contraintes et les préférences entre solutions. Puis, nous avons montré l’apport et la faisabilité de notre approche au travers d’une application en chémoinformatique portant sur la découverte de fragments moléculaires toxicophores. Les premiers résultats expérimentaux montrent l’ap-

port des contraintes souples de seuil pour mettre en évidence les structures chimiques les plus parlantes, comme par exemple les  $top_k$  motifs ou les JEPs.

Comme travaux futurs, nous souhaitons étudier l’apport des contraintes souples de seuil pour la résolution des problèmes de clustering sous contraintes [3], et pour le calcul des *skylines* qui constituent des points d’intérêt, non dominés par d’autres points, dans un espace [6].

## Références

- [1] J. Bajorath. Simulation of Sequential Screening Experiments Using Emerging Chemical Patterns. *Medicinal Chemistry*, 4 :80–90, 2008.
- [2] J. Bajorath and J. Auer. Emerging chemical patterns : A new methodology for molecular classification and compound selection. *J. of Chemical Information and Modeling*, 46 :2502–2514, 2006.
- [3] S. Basu, I. Davidson, and Kiri L. Wagstaff. *Constrained Clustering : Advances in Algorithms, Theory, and Applications*. Chapman & Hall, 2008.
- [4] Stefano Bistarelli and Francesco Bonchi. Interestingness is not a dichotomy : Introducing softness in constrained pattern mining. In *Knowledge Discovery in Databases (PKDD’05)*, volume 3721 of *LNCS*, pages 22–33. Springer, 2005.
- [5] Stefano Bistarelli and Francesco Bonchi. Soft constraint based pattern mining. *Data Knowl. Eng.*, 62(1) :118–137, 2007.
- [6] S. Börzsönyi, D. Kossmann, and K. Stocker. The skyline operator. In *ICDE 2001*, pages 421–430, 2001.
- [7] M. N. Garofalakis, R. Rastogi, and K. Shim. SPIRIT : Sequential pattern mining with regular expres-

N	intérêt	motif					émergence	fréquence	aromaticité	densité	SMILES	Condensée
$\lambda = 50$												
# Solutions=3												
1	253	24	35	87		$\infty$	47	66	88	cc ccc OP	ccc OP	
2	222	24	35	90		$\infty$	45	66	77	cc ccc OPO	ccc OPO	
3	222	24	35	105		$\infty$	45	66	77	cc ccc COP	ccc COP	
$\lambda = 60$												
# Solutions=457												
1	174	24	35	47	59	77	$\infty$	40	83	66	cc ccc cccc ccccc c1cccc1 c1cccc1 OP	
2	174	24	35	47	59	87	$\infty$	42	80	71	cc ccc cccc ccccc OP ccccc OP	
3	172	24	35	47	77	87	$\infty$	40	80	71	cc ccc cccc c1cccc1 OP c1cccc1 OP	
4	171	24	35	47	59	76	$\infty$	40	85	61	cc ccc cccc ccccc c1cccc1 OP c1cccc1 OP	
5	169	24	35	47	59	76	$\infty$	40	83	64	cc ccc cccc ccccc c1cccc1 OP c1cccc1 OP	
6	169	24	35	47	76	77	$\infty$	40	83	64	cc ccc cccc ccccc c1cccc1 OP c1cccc1 OP	
7	168	24	35	47	87		$\infty$	42	75	79	cc ccc cccc OP ccccc OP	
8	167	24	35	47	76	87	$\infty$	40	80	69	cc ccc cccc c1cccc1 OP c1cccc1 OP	
9	167	24	35	59	77	87	$\infty$	40	80	69	cc ccc cccc c1cccc1 OP c1cccc1 OP	
10	166	24	35	59	76	77	$\infty$	40	83	63	cc ccc cccc c1cccc1 OP c1cccc1 OP	
11	162	24	35	59	76	87	$\infty$	40	80	67	cc ccc cccc c1cccc1 OP c1cccc1 OP	
12	162	24	35	76	77	87	$\infty$	40	80	67	cc ccc cccc c1cccc1 OP c1cccc1 OP	
13	161	24	35	59	87		$\infty$	42	75	76	cc ccc cccc OP ccccc OP	
14	160	24	47	59	77	87	$\infty$	40	80	66	cc cccc c1cccc1 OP c1cccc1 OP	
15	159	24	35	77	87		$\infty$	40	83	60	cc ccc c1cccc1 c1cccc1 c1cccc1 OP	
16	159	24	47	59	76	77	$\infty$	40	75	76	cc cccc c1cccc1 c1cccc1 OP c1cccc1 OP	
17	157	24	35	59	76	77	$\infty$	38	83	60	cc ccc cccc c1cccc1 c1cccc1 OP c1cccc1 OP	
18	157	24	35	47	59	77	$\infty$	38	83	60	cc ccc cccc c1cccc1 c1cccc1 OPO c1cccc1 OPO	
19	156	24	35	47	76	77	$\infty$	38	83	59	cc ccc cccc c1cccc1 c1cccc1 COP c1cccc1 COP	
20	156	24	35	47	59	76	$\infty$	38	83	59	cc ccc cccc c1cccc1 c1cccc1 COP c1cccc1 COP	
21	156	24	35	47	76	77	$\infty$	38	83	59	cc ccc cccc c1cccc1 c1cccc1 OPO c1cccc1 OPO	
22	156	24	35	47	59	76	$\infty$	38	83	59	cc ccc cccc c1cccc1 c1cccc1 OPO c1cccc1 OPO	
23	155	24	47	76	77	87	$\infty$	40	80	64	cc cccc c1cccc1 OP c1cccc1 OP	
24	155	24	47	59	76	87	$\infty$	40	80	64	cc cccc c1cccc1 OP c1cccc1 OP	
25	155	24	35	47	59	105	$\infty$	40	80	64	cc ccc cccc COP cccc COP	
26	155	24	35	47	59	90	$\infty$	40	80	64	cc ccc cccc c1cccc1 OPO c1cccc1 OPO	

TABLE 5 –  $top_{25}$  Jumping Emerging Patterns extraits ( $\lambda=50$  et  $\lambda=60$ ).

- sion constraints. *In The VLDB Journal*, pages 223–234, 1999.
- [8] David J. Hand. Pattern detection and discovery. In *Pattern Detection and Discovery*, pages 1–12, 2002.
- [9] Y. Ke, J. Cheng, and J. Xu Yu. Top-k correlative graph mining. In *SDM*, pages 1038–1049, 2009.
- [10] M. Khiari, P. Boizumault, and B. Crémilleux. Constraint programming for mining n-ary patterns. In *CP'10*, volume 6308 of *LNCS*, pages 552–567. Springer, 2010.
- [11] M. Khiari, P. Boizumault, and B. Crémilleux. Combining CSP and constraint-based mining for pattern discovery. In *Advances in Knowledge Discovery and Management 2 (Post-EGC'10 Selected Papers)*, Studies in Computational Intelligence, Vol. 398, pages 85–104. Springer, March 2012.
- [12] K. Morik, J.F. Boulicaut, and A. Siebes, editors. *Local Pattern Detection, International Seminar, Dagstuhl Castle, Germany, April 12-16, 2004, Revised Selected Papers*, volume 3539 of *LNCS*. Springer, 2005.
- [13] S. Nijssen and T. Guns. Integrating constraint programming and itemset mining. In *European Conference, ECML PKDD 2010*, volume 6322 of *LNCS*, pages 467–482. Springer, 2010.
- [14] T. Petit, J.-C. Régin, and C. Bessière. Specific filtering algorithms for over-constrained problems. In *CP'01*, volume 2239 of *LNCS*, pages 451–463. Springer, 2001.
- [15] T. Petit, J.-C. Régin, C. Bessière, and J.-P. Puget. An original constraint based approach for solving over constrained problems. In *CP'2000*, volume 1894 of *LNCS*, pages 543–548. Springer, 2000.
- [16] G. Poezevara, B. Cuissart, B. Crémilleux, S. Lozano, M.-P. Halm-Lemeille, E. Lescot-Fontaine, A. Lepailleur, R. Bisell-Siders, S. Rault, and R. Bureau. Introduction of Jumping Fragments in Combination with QSARs for the Assessment of Classification in Ecotoxicology. *Journal of Chemical Information and Modeling(JCIM)*, pages 1330–1339, 2010.
- [17] L. De Raedt, T. Guns, and S. Nijssen. Constraint programming for itemset mining. In *KDD'08*, pages 204–212. ACM, 2008.
- [18] A. Soulet and B. Crémilleux. Optimizing constraint-based mining by automatically relaxing constraints. In *Proceedings of the 5th IEEE International Conference on Data Mining (ICDM 2005)*, pages 777–780, 2005.
- [19] Jianyong Wang, Jiawei Han, Ying Lu, and Petre Tzvetkov. Tfp : An efficient algorithm for mining top-k frequent closed itemsets. *IEEE Trans. Knowl. Data Eng.*, 17(5) :652–664, 2005.