PROGRAMMATION LINÉAIRE POUR L'EXTRACTION DE PATTERN SETS ET LEUR APPLICATION

Abdelkader Ouali

Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, Caen, FRANCE

Émail : abdelkader.ouali@unicaen.fr
Page web : https://ouali193.users.greyc.fr

GREYC
Laboratoire de recherche en sciences du numérique

Normandie Université

UNICAEN

ENSI CAEN
ÉCOLE PUBLIQUE D'INGÉNIEURS
CENTRE DE RECHERCHE

cnrs

Contexte - Représentation transactionnelle

- $\mathcal{I}$ ensemble de n littéraux distincts (items),
- $\mathcal{T}$ multi-ensemble (dataset) de m transactions t (motifs) t.q. $t \subseteq \mathcal{I}$,
- R relation binaire entre $\mathcal{T}$ et $\mathcal{I}$ t.q. $(t, i) \in R$ ssi $i \in t$.

- **Motif/Pattern** : décrit par un sous-ensemble d'items (littéraux)
  - $cover(\phi, \mathcal{T}) = \{t \in \mathcal{T} \mid \phi \subseteq t\}$
  - $freq(\phi, \mathcal{T}) = |cover(\phi)|$
  - $size(\phi, \mathcal{T}) = |\{i \in \mathcal{I} \mid i \in \phi\}|$
  - $area(\phi, \mathcal{T}) = size(\phi) \,.\, freq(\phi, \mathcal{T})$

Trans.	Items						
t_1	A	B		D			
t_2	A				E	F	
t_3	A				E		G
t_4	A				E		G
t_5		B			E		G
t_6		B			E		G
t_7			C		E		G
t_8			C		E		G
t_9			C		E			H
t_{10}			C		E			H
t_{11}			C			F	G	H

Contexte - Représentation transactionnelle

- $\mathcal{I}$ ensemble de n littéraux distincts (items),
- $\mathcal{T}$ multi-ensemble (dataset) de m transactions t (motifs) t.q. $t \subseteq \mathcal{I}$,
- R relation binaire entre $\mathcal{T}$ et $\mathcal{I}$ t.q. $(t, i) \in R$ ssi $i \in t$.

- **Concept formel :** un couple (I, T) tel que $I = int(T) \wedge T = ext(I)$

  ➠ **Concept formel $\equiv$ Motif fermé**

Motif fermé : Un motif ϕ est fermé ssi tous ses sur-ensembles stricts ont une fréquence strictement inférieure.

Trans.	Items						
t_1	A	B		D			
t_2	A			E	F		
t_3	A			E		G	
t_4	A			E		G	
t_5		B		E		G	
t_6		B		E		G	
t_7			C	E		G	
t_8			C	E		G	
t_9			C	E			H
t_{10}			C	E			H
t_{11}			C		F	G	H

Extraction de motifs

- Processus d'extraction sous contraintes portant sur des motifs individuels (motifs locaux)

  ➠ extraire la théorie $Th(\mathcal{L}, \mathcal{T}, q) = \{\phi \in \mathcal{L} \mid q(\phi, \mathcal{T}) \text{ est vrai}\}$, avec $\mathcal{L} = 2^{\mathcal{I}}$

- Problèmes de motifs fréquents

  - Motifs très fréquents ➠ motifs triviaux
  - Motifs peu fréquents ➠ bruit
  - Difficulté d'identifier des motifs intéressants

Motivation

- Une limitation bien connue en fouille de motifs : résultats incluant **un grand nombre** de motifs **non pertinents** ou **redondants**, difficultés à analyser !

- Un nouvel intérêt : vers des approches capables d'extraire d'ensembles de motifs plus pertinents

➥ Pattern sets ayant des **relations globales**

Extraction de pattern sets

Définition (Pattern set)

Soit $\mathcal{P}$ l'ensemble de tous les motifs locaux associés à la base transactionnelle $\mathcal{T}$.
Un pattern set $\Phi \subseteq \mathcal{P}$ est un ensemble de motifs.

Tâche d'extraction :

- Extraire la théorie $Th(\mathbf{L}, \mathcal{T}, p) = \{\Phi \in \mathbf{L} \mid p(\Phi, \mathcal{T}) \text{ est vrai}\}$,
  où $L = 2^{2^{\mathcal{I}}}$

Problèmes :

- Clustering conceptuel [Michalski et Stepp, 1983]
- Tiling [Geerts et al., 2004]

État de l'art

- Méthodes **heuristiques**
  ➥ Cluster Mining [Pensa et al., 2005], CDKMeans [Perkowitz & Etzioni, 1999], k-LTM[Geerts et al., 2004], . . .
  - **Solutions** en **temps raisonnable**, mais
  - aucune garantie sur l'**optimalité** des solutions
  - **qualité très variable** sur certains problèmes
  - flexibilité **limitée**
    ➠ nécessite de **revisiter** l'algorithme pour chaque nouvelle contrainte

- Approches **exactes** basées sur les motifs
  - Aposteriori [Zimmermann et al., 2007];
  - PPC : [Khiari et al., 2010], [Guns et al., 2013], [Chabert et al. 2017]. SAT : [Métivier et al.].

Passage à l'échelle : enjeu majeur pour les approches existantes

Approche en deux étapes

Approche offrant un **bon compromis** entre **flexibilité**, **optimalité** et **passage à l'échelle**

1. Étape d'extraction de motifs **locaux**
   - ➠ Utilisation d'extracteurs **dédiés** et **efficaces**

2. Modélisation et extraction du meilleur pattern set avec la **PLNE**
   - ➠ Formulation linéaire de **plusieurs contraintes** sur les pattern sets
   - ➠ Changement du **critère d'optimisation**

PLNE pour l'optimisation des pattern sets extraits.

Clustering conceptuel

Clustering conceptuel $\equiv$ ensemble de k concepts formels $\Phi = \{\phi_1, \ldots, \phi_k\}$, avec $\phi_j = (T_j, I_j)$, tel que $\{T_1, \ldots, T_k\}$ forme une partition de l'ensemble de transactions $\mathcal{T}$.

$$(Q_1) \begin{cases} k_{min} \leq k \leq k_{max} \wedge & \rightarrow\ k = size(\Phi) = |\Phi| \\ \bigcup_{i \in [1..k]} T_i = \mathcal{T} \wedge & \rightarrow\ \text{toutes les transactions sont couvertes} \\ \bigwedge_{i,j \in [1..k]} T_i \cap T_j = \emptyset \wedge & \rightarrow\ \text{sans aucun chevauchement entre clusters} \\ \bigwedge_{j \in [1..k]} \texttt{closed}(\phi_j) & \rightarrow\ \text{chaque concept formel est un motif fermé} \end{cases}$$

- $k_{min} = k_{max} = 3$

Trans.	Items							
t_1	A	B		D				
t_2	A				E	F		
t_3	A				E		G	
t_4	A				E		G	
t_5		B			E		G	
t_6		B			E		G	
t_7			C		E		G	
t_8			C		E		G	
t_9			C		E			H
t_{10}			C		E			H
t_{11}			C			F	G	H

Sol.	X_1	X_2	X_3
s_1	{C, F, G, H}	{E}	{A, B, D}
s_2	{B}	{C}	{A, E}
s_3	{A}	{C}	{B, E, G}

Clustering conceptuel

Clustering conceptuel $\equiv$ ensemble de k concepts formels $\Phi = \{\phi_1, \ldots, \phi_k\}$, avec $\phi_j = (T_j, I_j)$, tel que $\{T_1, \ldots, T_k\}$ forme une partition de l'ensemble de transactions $\mathcal{T}$.

$$(Q_1) \begin{cases} \max \sum_{\phi \in \Phi} size(\phi) & \\ k_{min} \leq k \leq k_{max} \wedge & \Rightarrow k = size(\Phi) = |\Phi| \\ \bigcup_{i \in [1..k]} T_i = \mathcal{T} \wedge & \Rightarrow \text{toutes les transactions sont couvertes} \\ \bigwedge_{i,j \in [1..k]} T_i \cap T_j = \emptyset \wedge & \Rightarrow \text{sans aucun chevauchement entre clusters} \\ \bigwedge_{j \in [1..k]} \texttt{closed}(\phi_j) & \Rightarrow \text{chaque concept formel est un motif fermé} \end{cases}$$

- $k_{min} = k_{max} = 3$

Trans.	Items							
t_1	A	B		D				
t_2	A				E	F		
t_3	A				E		G	
t_4	A				E		G	
t_5		B			E		G	
t_6		B			E		G	
t_7			C		E		G	
t_8			C		E		G	
t_9			C		E			H
t_{10}			C		E			H
t_{11}			C			F	G	H

Sol.	X_1	X_2	X_3
s_1	{C, F, G, H}	{E}	{A, B, D}
s_2	{B}	{C}	{A, E}
s_3	{A}	{C}	{B, E, G}

Clustering conceptuel

Clustering conceptuel $\equiv$ ensemble de k concepts formels $\Phi = \{\phi_1, \ldots, \phi_k\}$, avec $\phi_j = (T_j, I_j)$, tel que $\{T_1, \ldots, T_k\}$ forme une partition de l'ensemble de transactions $\mathcal{T}$.

$$(Q_1) \begin{cases} \max \sum_{\phi \in \Phi} size(\phi) \\ k_{min} \leq k \leq k_{max} \wedge \\ \bigcup_{i \in [1..k]} T_i = \mathcal{T} \wedge \\ \bigwedge_{i,j \in [1..k]} T_i \cap T_j = \emptyset \wedge \\ \wedge_{\phi \in \Phi} \ freq(\phi, \mathcal{T}) \geq 2 \\ \bigwedge_{j \in [1..k]} \texttt{closed}(\phi_j) \end{cases}$$

- $k = size(\Phi) = |\Phi|$
- toutes les transactions sont couvertes
- sans aucun chevauchement entre clusters
- chaque concept formel est un motif fermé

- $k_{min} = k_{max} = 3$

Trans.	Items							
t_1	A	B		D				
t_2	A				E	F		
t_3	A				E		G	
t_4	A				E		G	
t_5		B			E		G	
t_6		B			E		G	
t_7			C		E		G	
t_8			C		E		G	
t_9			C		E			H
t_{10}			C		E			H
t_{11}			C			F	G	H

Sol.	X_1	X_2	X_3
s_1	{C, F, G, H}	{E}	{A, B, D}
s_2	{B}	{C}	{A, E}
s_3	{A}	{C}	{B, E, G}

Clustering conceptuel en PLNE [IJCAI 2016]

$$(M1)\begin{cases} \texttt{Maximize } \sum_{p\in\mathcal{P}} v_p \,.\, x_p & \text{pour tout motif local } p,\ (x_p = 1) \text{ ssi } (p \in \Phi) \end{cases}$$

Clustering conceptuel en PLNE [IJCAI 2016]

$$(M1)\begin{cases} \texttt{Maximize } \sum_{p\in\mathcal{P}} v_p \,.\, x_p \\ (1) \sum_{p\in\mathcal{P}} a_{t,p} \,.\, x_p = 1 \quad \forall t \in \mathcal{T} \\ k = \sum_{p\in\mathcal{P}} x_p \\ k_{min} \leq k \leq k_{max} \\ k \in \mathbb{N}, \\ x_p \in \{0,1\}, \quad p \in \mathcal{P} \end{cases}$$

pour tout motif local p, $(x_p = 1)$ ssi $(p \in \Phi)$

	c_0	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_{10}	c_{11}	c_{12}	c_{13}	c_{14}	c_{15}	c_{16}	c_{17}
t_1	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0
t_2	1	0	1	1	0	1	0	0	0	0	0	0	0	0	1	0	1	0
t_3	1	0	1	0	1	0	0	0	0	0	0	0	0	0	1	1	0	1
t_4	1	0	1	0	1	0	0	0	0	0	0	0	0	0	1	1	0	1
t_5	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	1	0	1
t_6	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	1	0	1
t_7	0	0	0	0	0	0	0	1	1	1	0	0	1	0	1	1	0	1
t_8	0	0	0	0	0	0	0	1	1	1	0	0	1	0	1	1	0	1
t_9	0	0	0	0	0	0	0	1	1	0	1	0	0	1	1	0	0	0
t_{10}	0	0	0	0	0	0	0	1	1	0	1	0	0	1	1	0	0	0
t_{11}	0	0	0	0	0	0	0	1	0	0	0	1	1	1	0	0	1	1

TABLE – Matrice $(a_{t,p})$ de couverture par motif fermé.

Clustering conceptuel en PLNE [IJCAI 2016]

$$(M1)\begin{cases} \texttt{Maximize } \sum_{p\in\mathcal{P}} v_p \,.\, x_p \\ (1)\sum_{p\in\mathcal{P}} a_{t,p} \,.\, x_p = 1 \quad \forall t \in \mathcal{T} \\ k = \sum_{p\in\mathcal{P}} x_p \\ k_{min} \leq k \leq k_{max} \\ k \in \mathbb{N}, \\ x_p \in \{0,1\}, \quad p \in \mathcal{P} \end{cases}$$

Taille du modèle :

- Nombre de variables : $\Theta(|\mathcal{P}|)$
- Nombre de contraintes : $\Theta(|\mathcal{T}|)$

	c_0	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_{10}	c_{11}	c_{12}	c_{13}	c_{14}	c_{15}	c_{16}	c_{17}
t_1	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0
t_2	1	0	1	1	0	1	0	0	0	0	0	0	0	0	1	0	1	0
t_3	1	0	1	0	1	0	0	0	0	0	0	0	0	0	1	1	0	1
t_4	1	0	1	0	1	0	0	0	0	0	0	0	0	0	1	1	0	1
t_5	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	1	0	1
t_6	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	1	0	1
t_7	0	0	0	0	0	0	0	1	1	1	0	0	1	0	1	1	0	1
t_8	0	0	0	0	0	0	0	1	1	1	0	0	1	0	1	1	0	1
t_9	0	0	0	0	0	0	0	1	1	0	1	0	0	1	1	0	0	0
t_{10}	0	0	0	0	0	0	0	1	1	0	1	0	0	1	1	0	0	0
t_{11}	0	0	0	0	0	0	0	1	0	0	0	1	1	1	0	0	1	1

TABLE – Matrice $(a_{t,p})$ de couverture par motif fermé.

Clustering conceptuel : soft co-clustering

co-clustering $\equiv$ ensemble de concepts formels $\Phi = \{\phi_1, \ldots, \phi_k\}$, tel que Φ forme une partition sur l'ensemble de transactions $\mathcal{T}$ et l'ensemble d'items $\mathcal{I}$.

$$(Q_3) \begin{cases} k_{min} \leq k \leq k_{max} \wedge \\ \wedge_{\phi \in \Phi} \, closed(\phi) \wedge \\ cover(\Phi, \mathcal{T}) = |\mathcal{T}| \ \wedge cover(\Phi, \mathcal{I}) = |\mathcal{I}| \wedge \\ overlap(\Phi, \mathcal{T}) = 0 \ \wedge overlap(\Phi, \mathcal{I}) = 0 \end{cases}$$

Soft co-clustering $\equiv$ consiste à relâcher la couverture (δ_t, γ_i) et le non-chevauchement (δ_o, γ_o) sur l'ensemble des transactions et l'ensemble d'items

$$(Q_4) \begin{cases} k_{min} \leq k \leq k_{max} \wedge \\ \wedge_{\phi \in \Phi} \, closed(\phi) \wedge \\ cover(\Phi, \mathcal{T}) \lesssim \delta_t \ \wedge cover(\Phi, \mathcal{I}) \lesssim \gamma_i \wedge \\ overlap(\Phi, \mathcal{T}) \lesssim \delta_o \ \wedge overlap(\Phi, \mathcal{I}) \lesssim \gamma_o \end{cases}$$

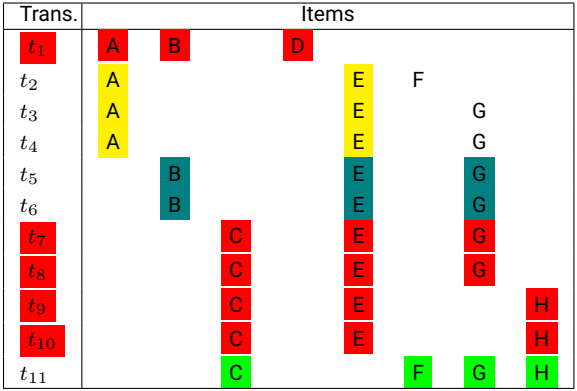
Exemple d'un soft co-clustering

- $k_{min} = 1, k_{max} = 3$
- Pour δ_t=6, δ_o=0, γ_i=7, γ_o=2
  - $s_r = [\{A, E\}, \{B, E, G\}, \{C, F, G, H\}]$

Trans.	Items							
t_1	A	B		D				
t_2	A				E	F		
t_3	A				E		G	
t_4	A				E		G	
t_5		B			E		G	
t_6		B			E		G	
t_7			C		E		G	
t_8			C		E		G	
t_9			C		E			H
t_{10}			C		E			H
t_{11}			C			F	G	H

Exemple d'un soft co-clustering

- $k_{min} = 1, k_{max} = 3$
- Pour δ_t=6, δ_o=0, γ_i=7, γ_o=2
  - $s_r = [\{A, E\}, \{B, E, G\}, \{C, F, G, H\}]$



Soft co-clustering en PLNE [IJCAI 2016]

$(M2)$ $\begin{cases} \texttt{Maximize } \sum_{p \in \mathcal{P}} v_p \,.\, x_p \end{cases}$

Soft co-clustering en PLNE [IJCAI 2016]

$(M2)$

Maximize $\sum_{p \in \mathcal{P}} v_p \,.\, x_p$

(1) $y_t \leq \sum_{p \in \mathcal{P}} a_{t,p} \,.\, x_p \leq \delta_o \,.\, y_t, \quad \forall t \in \mathcal{T}$

(2) $\sum_{t \in \mathcal{T}} y_t \geq \delta_t \quad y_t = 1$ ssi $\exists \phi \in \Phi$ t.q. $t \in cover(\phi, \mathcal{T})$

	ϕ_1	ϕ_2	ϕ_3	$\cdots$	ϕ_{18}
t_1	1	1	0	$\cdots$	0
t_2	1	0	1	$\cdots$	0
t_3	1	0	1	$\cdots$	1
t_4	1	0	1	$\cdots$	1
t_5	0	0	0	$\cdots$	1
t_6	0	0	0	$\cdots$	1
t_7	0	0	0	$\cdots$	1
t_8	0	0	0	$\cdots$	1
t_9	0	0	0	$\cdots$	0
t_{10}	0	0	0	$\cdots$	0
t_{11}	0	0	0	$\cdots$	1

(a) Matrice $(a_{t,p})$

Soft co-clustering en PLNE [IJCAI 2016]

$(M2)$

Maximize $\sum_{p\in\mathcal{P}} v_p \,.\, x_p$

(1) $y_t \leq \sum_{p\in\mathcal{P}} a_{t,p} \,.\, x_p \leq \delta_o \,.\, y_t, \quad \forall t \in \mathcal{T}$

(2) $\sum_{t\in\mathcal{T}} y_t \geq \delta_t \quad y_t = 1$ ssi $\exists \phi \in \Phi$ t.q. $t \in cover(\phi, \mathcal{T})$

(3) $z_i \leq \sum_{p\in\mathcal{P}} w_{i,p} \,.\, x_p \leq \gamma_o \,.\, z_i, \quad \forall i \in \mathcal{I}$

(4) $\sum_{i\in\mathcal{I}} z_i \geq \gamma_i \quad z_i = 1$ ssi $\exists \phi \in \Phi$ t.q. $i \in \phi$

	ϕ_1	ϕ_2	ϕ_3	$\cdots$	ϕ_{18}
t_1	1	1	0	$\cdots$	0
t_2	1	0	1	$\cdots$	0
t_3	1	0	1	$\cdots$	1
t_4	1	0	1	$\cdots$	1
t_5	0	0	0	$\cdots$	1
t_6	0	0	0	$\cdots$	1
t_7	0	0	0	$\cdots$	1
t_8	0	0	0	$\cdots$	1
t_9	0	0	0	$\cdots$	0
t_{10}	0	0	0	$\cdots$	0
t_{11}	0	0	0	$\cdots$	1

(a) Matrice $(a_{t,p})$

	ϕ_1	ϕ_2	ϕ_3	$\cdots$	ϕ_{18}
A	0	0	0	$\cdots$	1
B	0	0	0	$\cdots$	1
C	0	0	0	$\cdots$	0
D	0	0	0	$\cdots$	1
E	1	0	1	$\cdots$	0
F	0	0	0	$\cdots$	0
G	0	1	1	$\cdots$	0
H	0	0	0	$\cdots$	0

(b) Matrice $(w_{i,p})$

Soft co-clustering en PLNE [IJCAI 2016]

$$(M2)\begin{cases} \texttt{Maximize } \sum_{p\in\mathcal{P}} v_p \,.\, x_p \\ (1)\ y_t \le \sum_{p\in\mathcal{P}} a_{t,p} \,.\, x_p \le \delta_o \,.\, y_t, & \forall t \in \mathcal{T} \\ (2)\ \sum_{t\in\mathcal{T}} y_t \ge \delta_t & y_t = 1 \text{ ssi } \exists \phi \in \Phi \text{ t.q. } t \in cover(\phi, \mathcal{T}) \\ (3)\ z_i \le \sum_{p\in\mathcal{P}} w_{i,p} \,.\, x_p \le \gamma_o \,.\, z_i, & \forall i \in \mathcal{I} \\ (4)\ \sum_{i\in\mathcal{I}} z_i \ge \gamma_i & z_i = 1 \text{ ssi } \exists \phi \in \Phi \text{ t.q. } i \in \phi \\ k = \sum_{p\in\mathcal{P}} x_p \\ k_{min} \le k \le k_{max} \\ k \in \mathbb{N}, \\ x_p \in \{0,1\}, \quad p \in \mathcal{P} \\ y_t \in \{0,1\}, \quad t \in \mathcal{T} \\ z_i \in \{0,1\}, \quad i \in \mathcal{I} \end{cases}$$

	ϕ_1	ϕ_2	ϕ_3	$\cdots$	ϕ_{18}
t_1	1	1	0	$\cdots$	0
t_2	1	0	1	$\cdots$	0
t_3	1	0	1	$\cdots$	1
t_4	1	0	1	$\cdots$	1
t_5	0	0	0	$\cdots$	1
t_6	0	0	0	$\cdots$	1
t_7	0	0	0	$\cdots$	1
t_8	0	0	0	$\cdots$	1
t_9	0	0	0	$\cdots$	0
t_{10}	0	0	0	$\cdots$	0
t_{11}	0	0	0	$\cdots$	1

(a) Matrice ($a_{t,p}$)

	ϕ_1	ϕ_2	ϕ_3	$\cdots$	ϕ_{18}
A	0	0	0	$\cdots$	1
B	0	0	0	$\cdots$	1
C	0	0	0	$\cdots$	0
D	0	0	0	$\cdots$	1
E	1	0	1	$\cdots$	0
F	0	0	0	$\cdots$	0
G	0	1	1	$\cdots$	0
H	0	0	0	$\cdots$	0

(b) Matrice ($w_{i,p}$)

Soft co-clustering en PLNE [IJCAI 2016]

$$(M2)\begin{cases} \texttt{Maximize } \sum_{p\in\mathcal{P}} v_p \,.\, x_p \\ (1)\ y_t \le \sum_{p\in\mathcal{P}} a_{t,p} \,.\, x_p \le \delta_o \,.\, y_t, \quad \forall t \in \mathcal{T} \\ (2)\ \sum_{t\in\mathcal{T}} y_t \ge \delta_t \\ (3)\ z_i \le \sum_{p\in\mathcal{P}} w_{i,p} \,.\, x_p \le \gamma_o \,.\, z_i, \quad \forall i \in \mathcal{I} \\ (4)\ \sum_{i\in\mathcal{I}} z_i \ge \gamma_i \\ k = \sum_{p\in\mathcal{P}} x_p \\ k_{min} \le k \le k_{max} \\ k \in \mathbb{N}, \\ x_p \in \{0,1\}, \quad p \in \mathcal{P} \\ y_t \in \{0,1\}, \quad t \in \mathcal{T} \\ z_i \in \{0,1\}, \quad i \in \mathcal{I} \end{cases}$$

Taille du modèle :

- Nombre de variables : $\Theta(|\mathcal{P}| + |\mathcal{T}| + |\mathcal{I}|)$
- Nombre de contraintes : $\Theta(2 \,.\, |\mathcal{T}| + 2 \,.\, |\mathcal{I}|)$

	ϕ_1	ϕ_2	ϕ_3	$\cdots$	ϕ_{18}
t_1	1	1	0	$\cdots$	0
t_2	1	0	1	$\cdots$	0
t_3	1	0	1	$\cdots$	1
t_4	1	0	1	$\cdots$	1
t_5	0	0	0	$\cdots$	1
t_6	0	0	0	$\cdots$	1
t_7	0	0	0	$\cdots$	1
t_8	0	0	0	$\cdots$	1
t_9	0	0	0	$\cdots$	0
t_{10}	0	0	0	$\cdots$	0
t_{11}	0	0	0	$\cdots$	1

(a) Matrice ($a_{t,p}$)

	ϕ_1	ϕ_2	ϕ_3	$\cdots$	ϕ_{18}
A	0	0	0	$\cdots$	1
B	0	0	0	$\cdots$	1
C	0	0	0	$\cdots$	0
D	0	0	0	$\cdots$	1
E	1	0	1	$\cdots$	0
F	0	0	0	$\cdots$	0
G	0	1	1	$\cdots$	0
H	0	0	0	$\cdots$	0

(b) Matrice ($w_{i,p}$)

Clustering conceptuel équilibré [PAKDD 2018]

Trans.	Items							
t_1	A	B		D				
t_2	A				E	F		
t_3	A				E		G	
t_4	A				E		G	
t_5		B			E		G	
t_6		B			E		G	
t_7			C		E		G	
t_8			C		E		G	
t_9			C		E			H
t_{10}			C		E			H
t_{11}			C			F	G	H

Clustering conceptuel équilibré [PAKDD 2018]

- **Fonctions d'agrégation.** fonction d'utilité collective (CUF), les plus utilisées sont : MaxMin, MinDev et MaxSum.

  ➠ l'**effet de noyade** : (0, 1, 1, 1) et (1000, 1000, 1000, 0), pourtant très différents, mais impossible de distinguer par le MaxMin

- Approche qui trouve un clustering conceptuel **le plus équilibré**

  ➥ Si aucun intérêt à des **solutions extrêmes** favorisant un concepts au détriment des autres

  ➥ Modèle de préférence initial **neutre** qui permet d'aller progressivement avec les retours utilisateur

Clustering conceptuel équilibré [PAKDD 2018]

- Appliquer les principes de l'équité sur un opérateur d'agrégation (i.e. OWA)
  - **Symétrie**. $x \in \mathbb{R}_+^n$, pour toute permutation σ sur N, on a $(x_{\sigma(1)}, ..., x_{\sigma(n)}) \sim (x_1, ..., x_n)$
    e.g. (5, 3, 0) et (0, 3, 5).
  - **P-Monotonie.** $\forall x, y \in \mathbb{R}_+^n, x \succsim_P y \Rightarrow x \succsim_{\|} y$ and $x \succ_P y \Rightarrow x \succ_{\|} y$.
    e.g. x = (**5**, 5, 1) $\succsim_P$ et $\succsim_{\|}$ y = (**3**, 5, 1)
  - **Principe de transfert.** Let $x \in \mathbb{R}_+^n$ and $x_i > x_j$ for some $i, j \in N$. Let e^z be a vector such that $\forall i \neq z, e_i^z = 0$ and $e_z^z = 1$. For all ϵ where $0 < \epsilon \leq \frac{x_i - x_j}{2}$, we get $x - \epsilon e^i + \epsilon e^j \succsim_{\|} x$.
    e.g. y = (**9**, 10, **9**, 10) $\succsim_{\|}$ x = (**11**, 10, **7**, 10), $\epsilon = 2$

Clustering conceptuel équilibré [PAKDD 2018]

- **Infinite Generalized Lorenz dominance** $\succsim_L^\infty$ pour assurer les principes de l'équité.
- **Représentation numérique directe** de $\succsim_L^\infty$, cette représentation est donnée par la propriété suivante : $\forall x, y \in \mathbb{R}_+^n, x \succsim_L^\infty y \Leftrightarrow W(x) > W(y)$

Ordered Weighted Averaging aggregation

C'est une famille d'agrégateurs introduite par Yager caractérisée par :

$$W(x) = \sum_{k=1}^{n} w_k x_{(k)}$$

où $w = (w_1, \ldots, w_n) \in [0,1]^n$, $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ et $w_k = sin(\frac{(n+1-k)\pi}{2n+1})$ coefficients de Golden & Perny pour l'agrégateur OWA afin que la fonction soit Schur-convex.

Modèle OWA-PLNE pour clustering conceptuel équilibré.

- Un clustering conceptuel est évalué sur un **vecteur**.
- Chaque concept ayant sa **propre utilité** correspondant à une mesure spécifique (e.g. la fréquence).

$$
\begin{aligned}
&\texttt{Max} \sum_{p=1}^{|\mathcal{P}|} \omega_p \,.\, r_p \\
&\text{s.t.} \begin{cases}
\text{Clustering.} \begin{cases}
\text{(C1)} & \sum_{p=1}^{|\mathcal{P}|} a_{t,p} \,.\, x_p = 1, & \forall t \in \mathcal{T} \\
\text{(C2)} & k_{min} \leq \sum_{p=1}^{|\mathcal{P}|} x_p \leq k_{max}
\end{cases} \\
\text{OWA sorting.} \begin{cases}
\text{(O1)} & r_p - (v_i \,.\, x_i) \leq M z_{p,i}, & \forall\, i, p = 1, ..., |\mathcal{P}| \\
\text{(O2)} & \sum_{i=1}^{|\mathcal{P}|} z_{p,i} \leq p - 1, & \forall\, p = 1, ..., |\mathcal{P}|
\end{cases} \\
x_p \in \{0,1\},\ r_p \in \mathbb{R}_+, \quad \forall\, p = 1, ..., |\mathcal{P}| \\
z_{p,i} \in \{0,1\}, \quad \forall\, i, p = 1, ..., |\mathcal{P}|
\end{cases}
\end{aligned}
$$

Modèle OWA-PLNE amélioré

$$\texttt{Max} \sum_{c=1}^{|\mathcal{C}|} \omega_c \,.\, (v_c^{\uparrow} \,.\, x_c^{\uparrow})$$

$$\text{s.t.} \begin{cases} \text{(C1)}, \quad \text{(C2)} \\ x_c \in \{0,1\}, \\ \quad \forall\, c = 1, ..., |\mathcal{C}| \end{cases}$$

- **Contraintes de tri.** Les valeurs d'utilité sont connues à l'avance. Le tri est **effectué immédiatement après** l'extraction de motifs fermés.

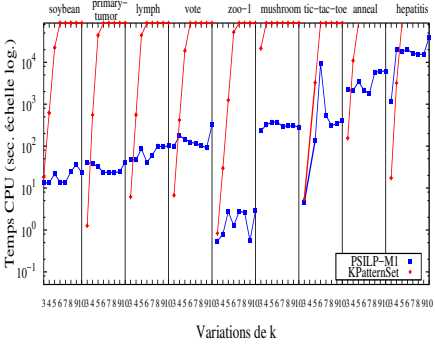
Jeux de données

datasets	#transactions	#items	densité(%)	#motifs locaux
Soybean	630	50	32	31,759
Primary-tumor	336	31	48	87,230
Lymph	148	68	40	154,220
Vote	435	48	33	227,031
tic-tac-toe	958	27	33	42,711
Mushroom	**8124**	119	18	221,524
Zoo-1	101	36	44	4,567
Hepatitis	137	68	50	**3,788,341**
Anneal	812	93	45	**1,805,193**

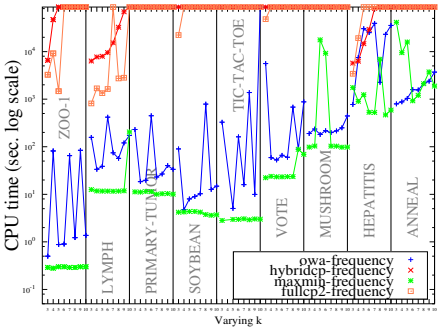
TABLE – Caractéristiques des différents jeux de données.

Passage à l'échelle et temps CPU (2)

- Solutions optimales sur des bases de données plus difficiles.



(a) Clustering conceptuel



(b) Clustering conceptuel équilibré

Qualité des clusterings : (similarité)

- Distance de Jaccard :

$$s : \mathcal{T} \times \mathcal{T} \mapsto [0,1], \quad s(t,t') = \frac{|t \cap t'|}{|t \cup t'|}$$

- Similarité intra-cluster :

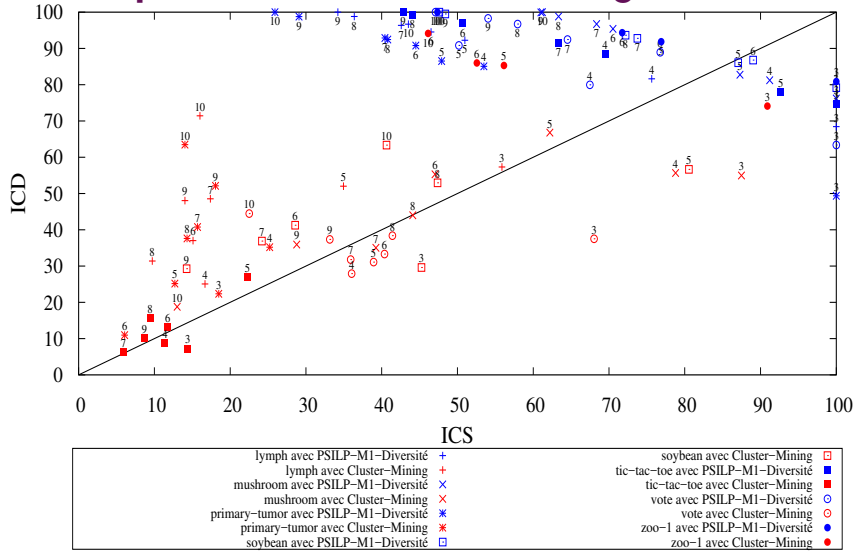
$$ICS(P_1,...,P_k) = \frac{1}{2} \sum_{1 \leq i \leq k} \Big(\sum_{t,t' \in P_i} s(t,t') \Big)$$

- Dissimilarité inter-clusters :

$$ICD(P_1,...,P_k) = \sum_{1 \leq i < j \leq k} \Big(\sum_{t \in P_i, t' \in P_j} (1 - s(t,t')) \Big)$$

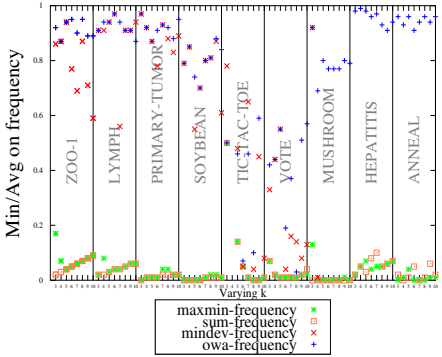
- Choix du **meilleur clustering** pour **Cluster Mining** :
  1. Exécuter chaque méthode 100 fois,
  2. Regrouper les résultats en *classes d'équivalence*,
  3. Choisir la plus grande classe d'équivalence et retenir son représentant.

ICS et ICD : comparaison avec Cluster Mining

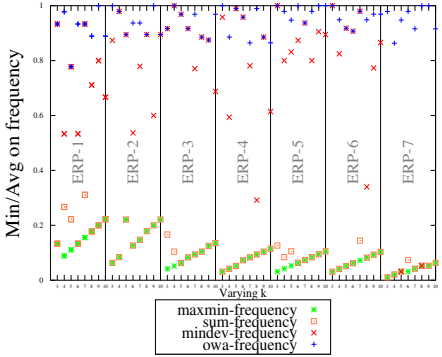


Qualité des clusterings équilibrés des modèles PLNE.

- Ratio entre la fréquence minimale et la moyenne (Min/Avg).



(c) Évaluation (Min/Avg) sur UCI datasets.



(d) Évaluation (Min/Avg) sur ERP datasets.

Autres contraintes sur les patterns sets [PAKDD 2017]

$$(M3)\begin{cases} \texttt{Maximize } \sum_{t\in\mathcal{D}, i\in\mathcal{I}} \mathbf{q_{t,i}} \\ \mathbf{q_{t,i}} \leq \sum_{p\in\mathcal{P}} cq^{p}_{t,i} x_p \leq |\mathcal{P}| \,.\, \mathbf{q_{t,i}}, \forall t \in \mathcal{T},\ \forall i \in \mathcal{I} \\ \sum_{t\in\mathcal{D}, i\in\mathcal{I}} \mathbf{q_{t,i}} \leq \theta_a \\ x_p \in \{0,1\}, \quad p \in \mathcal{P} \\ q_{i,t} \in \{0,1\}, \quad i \in \mathcal{I}, t \in \mathcal{T} \end{cases}$$

- Contrainte de l'**aire** : soit Ψ un pattern set, l'aire de Ψ est donnée par : $area(\Psi) = |\bigcup_{\phi\in\Phi} area(\psi)|$, où $area(\psi) = \{(t,i) | t \in cover(area), i \in \psi\}$

	A	B	C	D	E	F	G	H
t_1	0	0	0	0	0	0	0	0
t_2	0	0	0	0	1	0	0	0
t_3	0	0	0	0	1	0	0	0
t_4	0	0	0	0	1	0	0	0
t_5	0	0	0	0	1	0	0	0
t_6	0	0	0	0	1	0	0	0
t_7	0	0	0	0	1	0	0	0
t_8	0	0	0	0	1	0	0	0
t_9	0	0	0	0	1	0	0	0
t_{10}	0	0	0	0	1	0	0	0
t_{11}	0	0	0	0	0	0	0	0

TABLE – Matrice de couverture $cq^{\{E\}}_{t,i}$ du motif $\{E\}$.

Autres contraintes sur les patterns sets [PAKDD 2017]

- Contrainte de **redondance** : soit Ψ un pattern set, les transactions redondantes de Ψ sont données par :
  $red(\Phi, \mathcal{T}) = |\{t \in \mathcal{T} | \exists (\phi, \psi) \in \Phi, t \in overlap(\phi, \psi, \mathcal{T})\}|$

$$(M3) \begin{cases} \texttt{Minimize } \sum_{t \in \mathcal{T}} \mathbf{u_t} \\ 2\mathbf{u_t} \leq \sum_{p \in \mathcal{P}} a_{t,p} x_p \leq |\mathcal{P}| \,.\, \mathbf{u_t}, \forall t \in \mathcal{T} \\ \sum_{t \in \mathcal{T}} \mathbf{u_t} \geq \theta_r \\ x_p \in \{0,1\}, \quad p \in \mathcal{P} \\ u_t \in \{0,1\}, \quad t \in \mathcal{T} \end{cases}$$

Soit le pattern set suivant :

- motif 3 : $\{E, G\}$ couvrant les transactions $\{\mathbf{t_3}, \mathbf{t_4}, t_5, t_6, t_7, t_8\}$
- motif 8 : $\{A\}$ couvrant les transactions $\{t_1, t_2, \mathbf{t_3}, \mathbf{t_4}\}$
- motif 13 : $\{C, H\}$ couvrant les transactions $\{t_9, t_{10}, t_{11}\}$

Autres contraintes sur les patterns sets [PAKDD 2017]

- Contrainte de **représentativité** : soit Ψ un pattern set, et $\mathcal{T}_i$ une base partielle de transactions de $\mathcal{T}$, la représentativité est donnée par : $rep(\Phi, \mathcal{T}_i, \mathcal{T}) = freq(\Phi, \mathcal{T}_i)/freq(\Phi, \mathcal{T})$

$$(M3) \begin{cases} \texttt{Minimize } \sum_{p\in\mathcal{P}} v_p x_p \\ \mathbf{y_t} \leq \sum_{p\in\mathcal{P}} a_{t,p} \cdot x_p \leq |\mathcal{P}| \cdot \mathbf{y_t}, & \forall t \in \mathcal{T}. \\ \mathbf{y'_t} \leq \sum_{p\in\mathcal{P}} a_{t,p} \cdot x_p \leq |\mathcal{P}| \cdot \mathbf{y'_t}, & \forall t \in \mathcal{T}_i. \\ \sum_{t\in\mathcal{T}_i} y'_t \leqslant \theta_{rep} \times \sum_{t\in\mathcal{T}} y_t. \\ x_p \in \{0,1\}, \quad p \in \mathcal{P} \\ y_t \in \{0,1\}, \quad t \in \mathcal{T} \\ y'_t \in \{0,1\}, \quad t \in \mathcal{T}_i \end{cases}$$

Soit le pattern set $\Psi = \{\psi_1, \psi_2\}$, soient $\mathcal{T}^+ = \{t_1, t_2, \ldots, t_5\}$:

- $\psi_1 = \{A, E\}$ couvrant les transactions $\{\mathbf{t_2}, \mathbf{t_3}, \mathbf{t_4}\}$
- $\psi_2 = \{B, E, G\}$ couvrant les transactions $\{\mathbf{t_5}, t_6\}$

$rep(\Psi, \mathcal{T}^+, \mathcal{T}) = 4/11 = 0.36$

Autres contraintes sur les patterns sets [PAKDD 2017]

- Contrainte de **spécialisation** : Soit Ψ un pattern set donné, Φ le pattern set recherché, généralisation est donne comme suit : $\Psi \preceq \Phi$ ssi $cover(\Phi, \mathcal{T}) \subseteq cover(\Psi, \mathcal{T})$

$$(M3) \begin{cases} \texttt{Optimize } \sum_{p \in \mathcal{P}} v_p x_p \\ \sum_{\{p:\, \mathbf{p} \in \mathcal{P} \mid \exists \psi \in \Psi,\ \psi \preceq \mathbf{p}\}} \mathbf{x_p} \geq \sum_{p \in \mathcal{P}} x_p \\ x_p \in \{0,1\}, \quad p \in \mathcal{P} \end{cases}$$

Soit Ψ un pattern set connu constitué des motifs :

- $\{A\}$ couvrant les transactions $\{t_1, \mathbf{t_2}, \mathbf{t_3}, \mathbf{t_4}\}$.
- $\{C\}$ couvrant les transactions $\{\mathbf{t_7}, \mathbf{t_8}, t_9, t_{10}, \mathbf{t_{11}}\}$.
- $\{C, H\}$ couvrant les transactions $\{\mathbf{t_9}, \mathbf{t_{10}}, t_{11}\}$.

Le pattern set Φ suivant est une spécialisation de Ψ

- $\{A, E\}$ couvrant les transactions $\{t_2, t_3, t_4\}$.
- $\{C, G\}$ couvrant les transactions $\{t_7, t_8, t_{11}\}$.
- $\{C, E, H\}$ couvrant les transactions $\{t_9, t_{10}\}$.

Autres contraintes sur les patterns sets [PAKDD 2017]

- Contrainte de **généralisation** : Soit Ψ un pattern set donné, Φ le pattern set recherché, généralisation est donne comme suit : $\Phi \preceq \Psi$ ssi $cover(\Psi, \mathcal{T}) \subseteq cover(\Phi, \mathcal{T})$

$$(M3) \begin{cases} \texttt{Optimize } \sum_{p \in \mathcal{P}} v_p x_p \\ \sum_{\{p:\, \forall \mathbf{p} \in \mathcal{P} \mid \mathbf{p} \preceq \psi\}} \mathbf{x_p} \geq 1, \quad \forall \psi \in \Psi. \\ x_p \in \{0,1\}, \quad p \in \mathcal{P} \end{cases}$$

Soit Ψ un pattern set connu constitué des motifs :

- $\{A, E\}$ couvrant les transactions $\{t_2, t_3, t_4\}$.
- $\{C, G\}$ couvrant les transactions $\{t_7, t_8, t_{11}\}$.
- $\{C, E, H\}$ couvrant les transactions $\{t_9, t_{10}\}$.

Le pattern set Φ suivant est une généralisation de Ψ

- $\{A\}$ couvrant les transactions $\{t_1, \mathbf{t_2}, \mathbf{t_3}, \mathbf{t_4}\}$.
- $\{C\}$ couvrant les transactions $\{\mathbf{t_7}, \mathbf{t_8}, \mathbf{t_9}, \mathbf{t_{10}}, \mathbf{t_{11}}\}$.

Application en chémoinformatique [1]

- Découvrir des **alertes structurelles** dans un ensemble de données biologiques et chimiques
- Relations qui relient les **structures chimiques** et **les activités toxicologiques**
- Caractéristiques clés d'une molécule qui sont nécessaires pour interagir avec un système biologique et initier une voie toxicologique
- la **fréquence** d'une sous-structure chimique dans un ensemble de données
  ➠ souvent au cœur du processus de définition de la **pertinence toxicologique**

1. Métivier et al. Discovering Structural Alerts for Mutagenicity Using Stable Emerging Molecular Patterns, 2015

Application en chémoinformatique [2]

- Contrainte émergente capture les caractéristiques qui différencient deux classes de données

- [Métivier et al. 2005] : calcule les conjonctions de fragments moléculaires

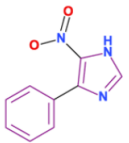
➥ fréquences d'occurrence dans un ensemble de données sont suffisamment **discriminantes**
➥ entre différents **sous-groupes de molécules** (par exemple, mutagènes et non mutagènes) pour être intéressantes.

2. Métivier et al. Discovering Structural Alerts for Mutagenicity Using Stable Emerging Molecular Patterns, 2015

Fragment moléculaire

Graphe moléculaire

Un graphe moléculaire est défini par $G = (V, E)$, où V est l'ensemble de nœuds représentant **les atomes**, et E est l'ensemble d'arêtes représentant les **liaisons chimiques** (interaction entre les atomes).



- Un **fragment moléculaire** représente **une partie connectée** d'une molécule **sous-graphe**
- Un fragment apparaît dans une molécule s'il y une imbrication qui satisfait :
  - la **structure relationnelle** du fragment
  - les **schémas d'étiquetage** des arêtes et des atomes
- La couverture d'un fragment moléculaire f désigne l'ensemble de molécules dans lesquelles f apparaît
- un fragment f est fermé en M s'il n'y a pas de fragment qui contient f et ayant le même couverture de f,

➠ **Objectif :** découvrir des **patterns moléculaires** qui correspondent à des **alertes structurelles potentielles**.

Pattern moléculaire

- Un **pattern moléculaire** p est un ensemble (**conjonction**) de fragments moléculaires
- la **longueur** d'un pattern moléculaire $size(p)$ désigne le **nombre de fragments** qu'il contient.
- Un pattern moléculaire apparaît dans une molécule si **chacun de ses fragments** apparaît dans la molécule.
- la couverture $cover(p)$ désigne l'ensemble de molécules dans lesquelles le pattern moléculaire p apparaît,
  ➠ $freq(p) = |cover(p)|$

Soit deux patterns moléculaires p et q, $p \subset q$ si chaque fragment de p est contenu dans un fragment de q.

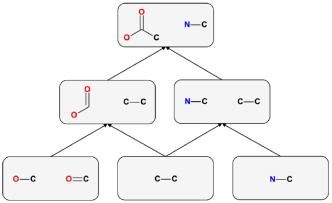


FIGURE – Exemple d'un diagramme de Hasse d'inclusion des patterns, [Métivier et al. 2015]

Pattern moléculaire fermé

- Un motif moléculaire p est **fermé** s'il n'y a pas un autre pattern moléculaire contenant p couvrant les mêmes molécules de p

- Tout fragment d'un motif fermé p est élagué s'il s'agit d'un sous-fragment d'un autre fragment de p;
  ➥ le motif résultant est nommé motif moléculaire fermé élagué.
- Conduire à l'information chimiquement la **plus interprétable**

- Chercher des **changements structurels** entre **différents groupes de molécules**
  - Frequent emerging molecular closed patterns
  - Mesure du taux de croissance
  - Mesure de stabilité

Modèles PLNE sur les patterns moléculaires (en cours)

- Extraire **patterns moléculaires fermés** candidats comme entrée pour l'approche en PLNE.

- **Contraintes linéaires** sur les patterns moléculaires
  - ➠ contraintes de **spécialisation** pour chercher des fragments informatif
  - ➠ contraintes de **représentativité** entre plusieurs sous-groupe de molécules

- Mesures comme le **taux de croissance** ou la **stabilité** comme critère d'optimisation

Conclusion & perspectives

- **Approches déclaratives :**
  - Permettant un **meilleur passage à l'échelle**
  - Extrayant des pattern sets de **meilleure qualité**
  - **Flexibilité**

- **Chémoinformatique :**
  - Ensembles de patterns moléculaires
  - Sélection de pharmacophores

- **Perspectives :**
  - Approches **interactives**

Merci !

Questions ?