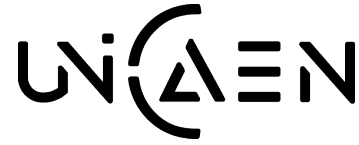
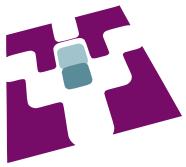


Contraintes globales pour la fouille de motifs séquentiels

A. Kemmar, Y. Lebbah, S. Loudni, P. Boizumault, T. Charnois

Journée CAVIAR

Montpelier, le 19 juin 2018



Plan de la présentation

1. Contexte applicatif

2. Préliminaires
   - Les motifs ensemblistes
   - Les motifs séquentiels (MS)
   - L'extraction de MS fréquents

3. Contributions
   - Une contrainte globale pour l'extraction de MS fréquents
   - Une contrainte globale pour l'extraction de MS sous contrainte de gap

4. Perspectives

Application : Analyse de textes biomédicaux

Maladies Rares : maladies affectant moins d'une personne sur 2 000

- Entre 6 000 et 8 000 MR recensées en Europe
  $\rightarrow$ **Orphanet** : Portail contenant une base de données internationale de connaissances sur les MR

Résumés d'articles de PubMed (17500 séquences)

- Collection d'articles synthétiques sur les MR
- Phrases contenant au moins un nom de gène et une MR

Num.	Phrase
1	We conclude that **VCP** is essential for maturation of ubiquitin containing autophagosomes and that defect in this function may contribute to **IBMPFD** pathogenesis.
2	**Osteogenesis imperfecta** is normally caused by an autosomal dominant mutation in the type I collagen genes **COL1A1** and **COL1A2**.

Application : Analyse de textes biomédicaux

Maladies Rares : maladies affectant moins d'une personne sur 2 000

- Entre 6 000 et 8 000 MR recensées en Europe
  $\rightarrow$ **Orphanet** : Portail contenant une base de données internationale de connaissances sur les MR

Résumés d'articles de PubMed (17500 séquences)

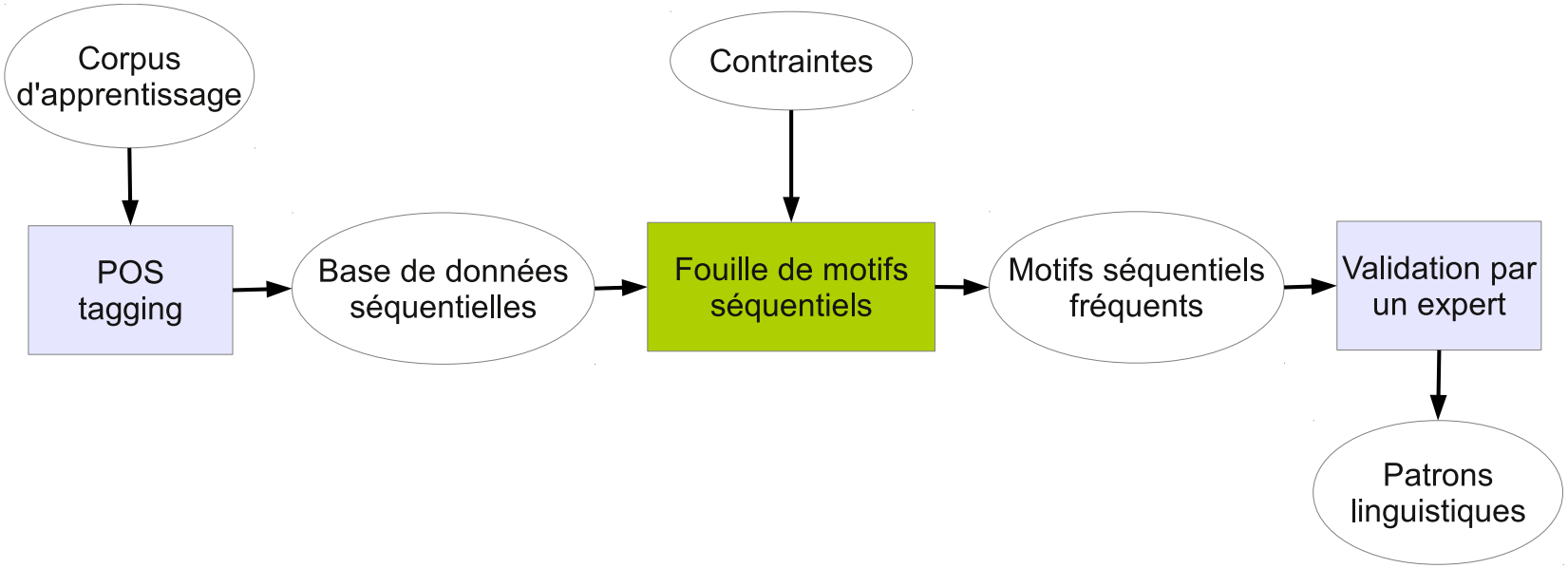
- Collection d'articles synthétiques sur les MR
- Phrases contenant au moins un nom de gène et une MR

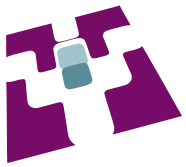
Num.	Phrase
1	We conclude that **VCP** is essential for maturation of ubiquitin containing autophagosomes and that defect in this function may contribute to **IBMPFD** pathogenesis.
2	**Osteogenesis imperfecta** is normally caused by an autosomal dominant mutation in the type I collagen genes **COL1A1** and **COL1A2**.

Application : Analyse de textes biomédicaux

- Problématique :
  - La veille nécessaire sur la parution de nouveaux articles dans la littérature est chronophage
  - La relecture de ceux-ci étaient des tâches réalisées manuellement
- La découverte de connaissances liées aux maladies rares à partir de textes est donc un enjeu particulièrement important
- Utilisation de la fouille de données pour extraire des relations entre gènes et MR

Vue générale de l'approche





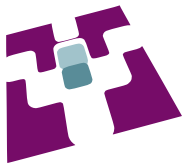
Contexte : Fouille de données

Définition

La fouille de données est une étape dans le processus d'extraction de connaissances, qui consiste à découvrir des informations nouvelles dans les BD. Le cœur du processus est la recherche de régularités [Agrawal 93].

Quelles techniques de fouilles de données

- l'extraction de motifs ensemblistes
- l'extraction de motifs séquentiels
- la fouille d'arbres
- la fouille de graphes
- etc.



Rappel du sommaire

1. Contexte applicatif
2. Préliminaires
   - Les motifs ensemblistes
   - Les motifs séquentiels (MS)
   - L'extraction de MS fréquents
3. Contributions
   - Une contrainte globale pour l'extraction de MS fréquents
   - Une contrainte globale pour l'extraction de MS sous contrainte de gap
4. Perspectives

Les motifs ensemblistes

Achat 1

Achat 2

Achat 3

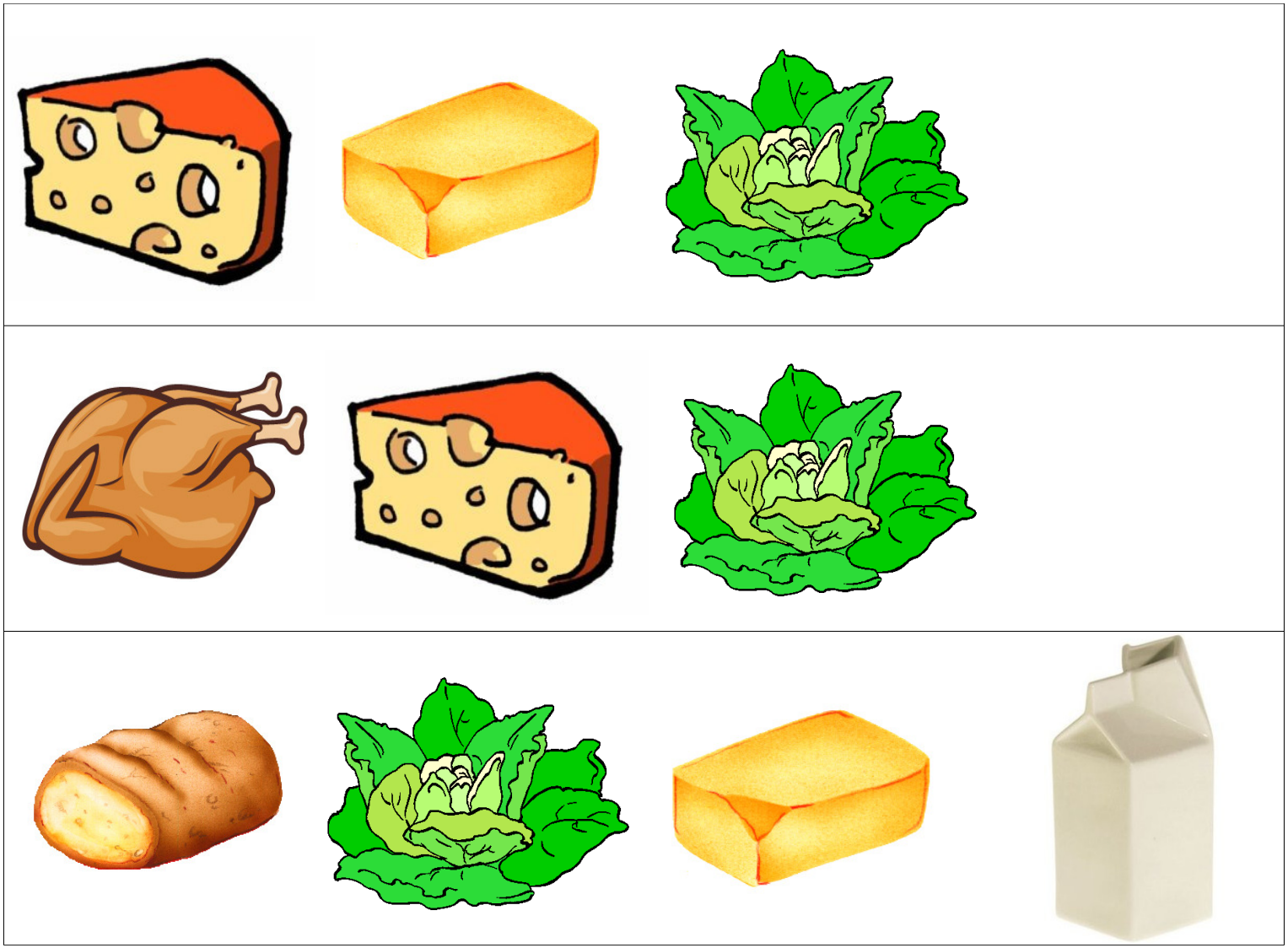
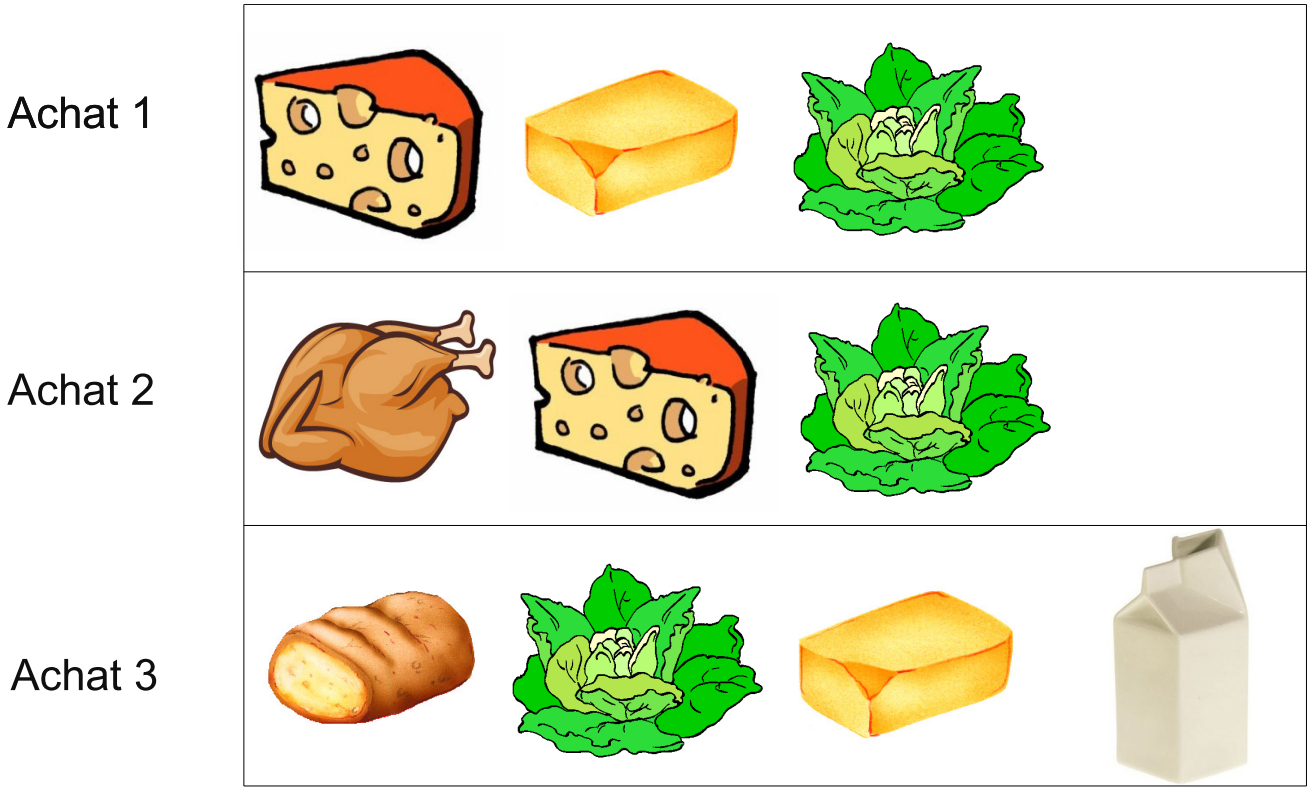


FIGURE – Exemple de données : les achats réalisés lors de courses

Les motifs ensemblistes



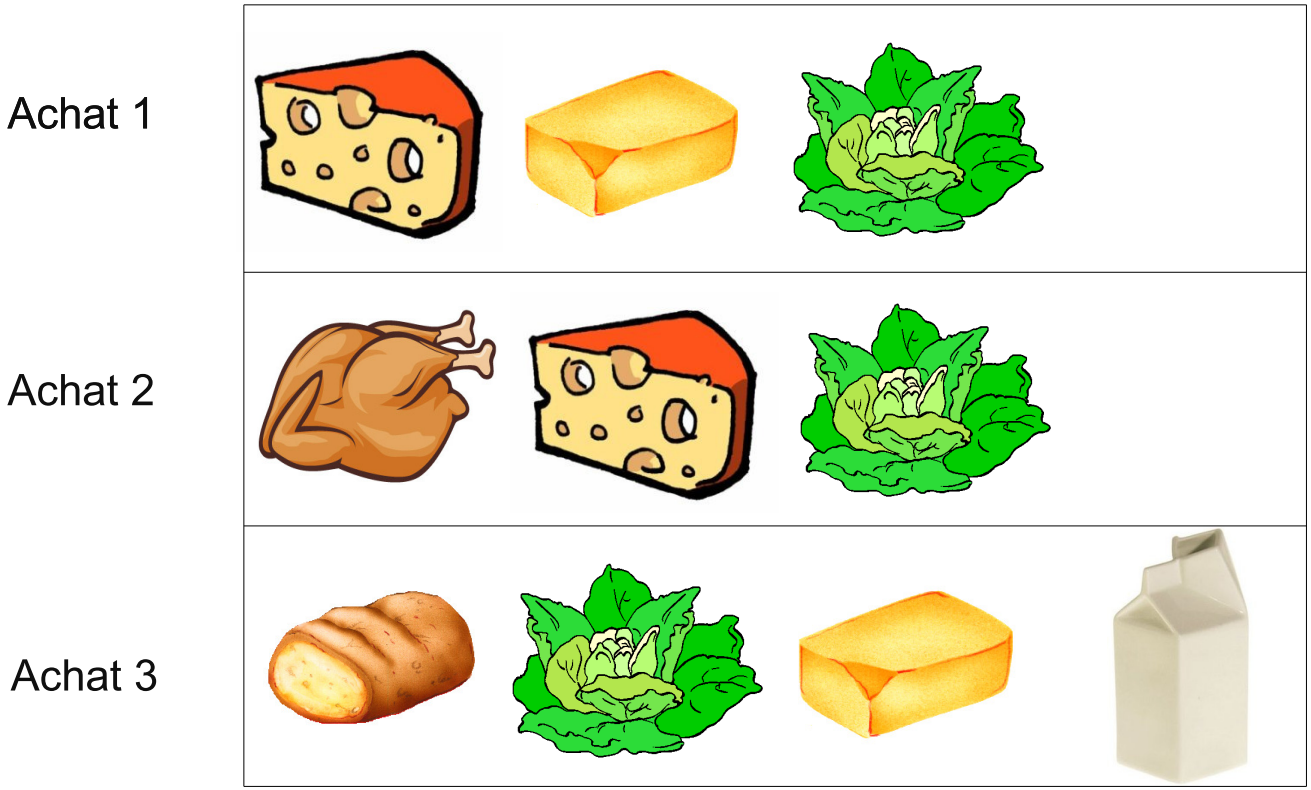
Achats fréquents ?

- salade (3 achats),
- salade-fromage (2 achats),
- salade-beurre (2 achats)

Extraction de règles ?

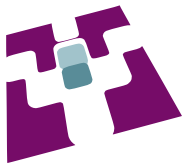
- fromage $\Rightarrow$ salade
- beurre $\Rightarrow$ salade
- beurre, salade $\Rightarrow$ lait (50% des cas)

Les motifs ensemblistes



Définitions

- Les **produits** sont appelés des items
- Les **achats** sont appelés des itemsets
- L'ensemble "fromage, salade" est un motif

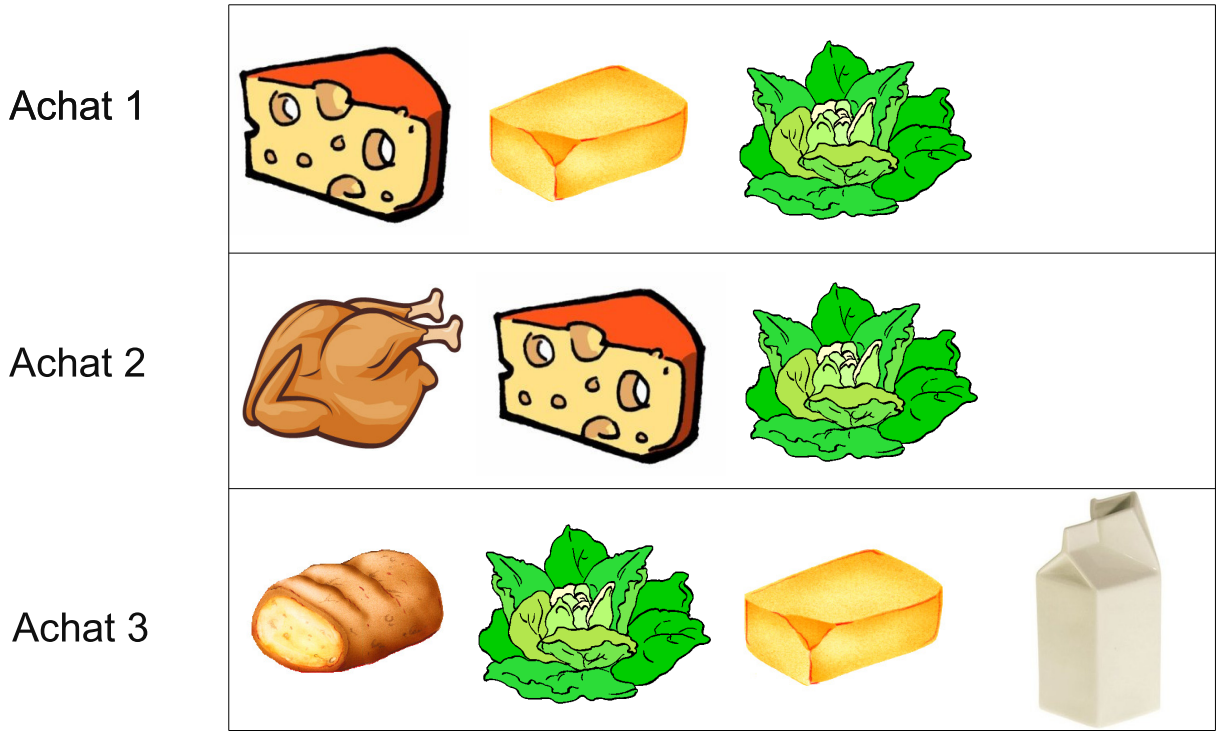


Rappel du sommaire

1. Contexte applicatif
2. Préliminaires
   - Les motifs ensemblistes
   - Les motifs séquentiels (MS)
   - L'extraction de MS fréquents
3. Contributions
   - Une contrainte globale pour l'extraction de MS fréquents
   - Une contrainte globale pour l'extraction de MS sous contrainte de gap
4. Perspectives

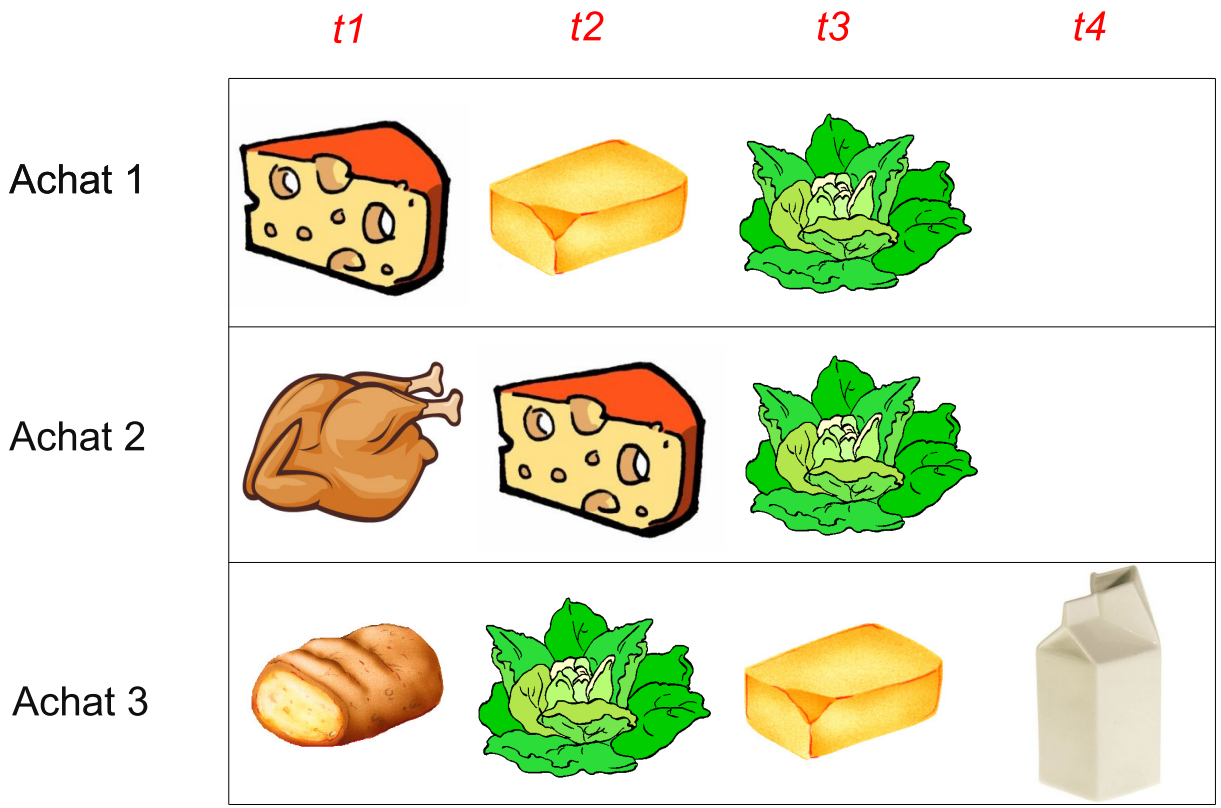
Des ensembles aux séquences

Extraction de MS : *extension de l'extraction de motifs ensemblistes avec prise en compte de la temporalité dans les données à étudier*

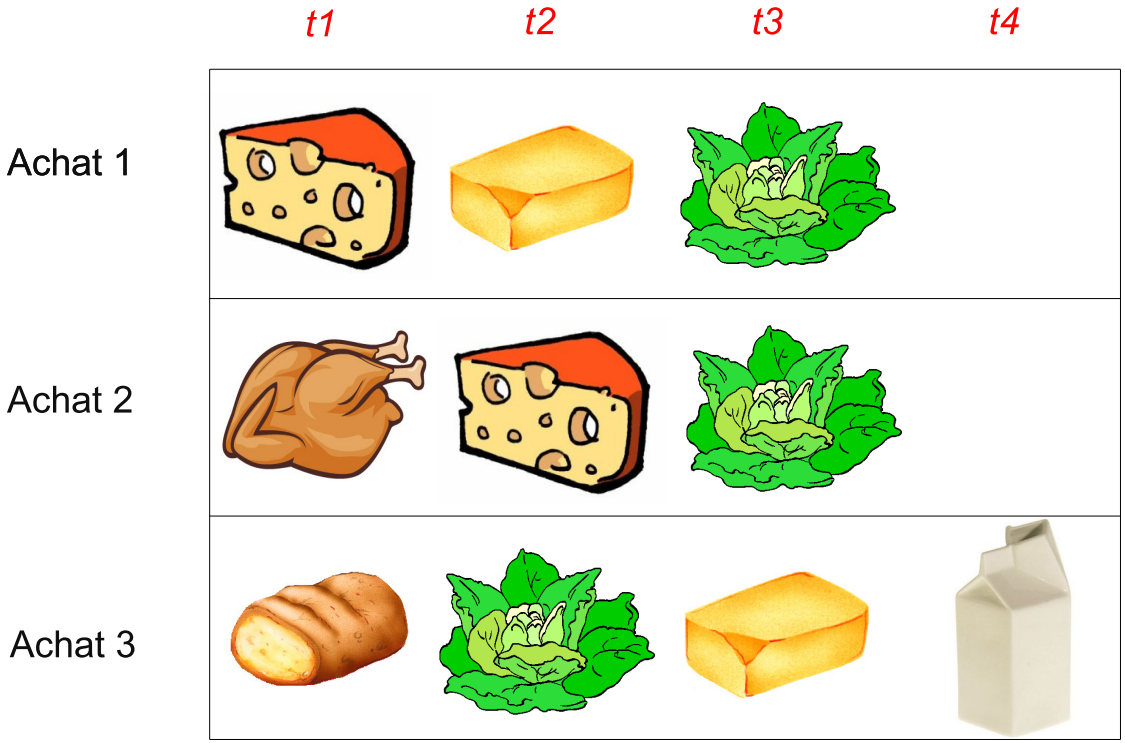


Exemple 1 : Prise en compte de l'ordre d'achat

Les produits sont achetés séquentiellement



Les motifs séquentiels

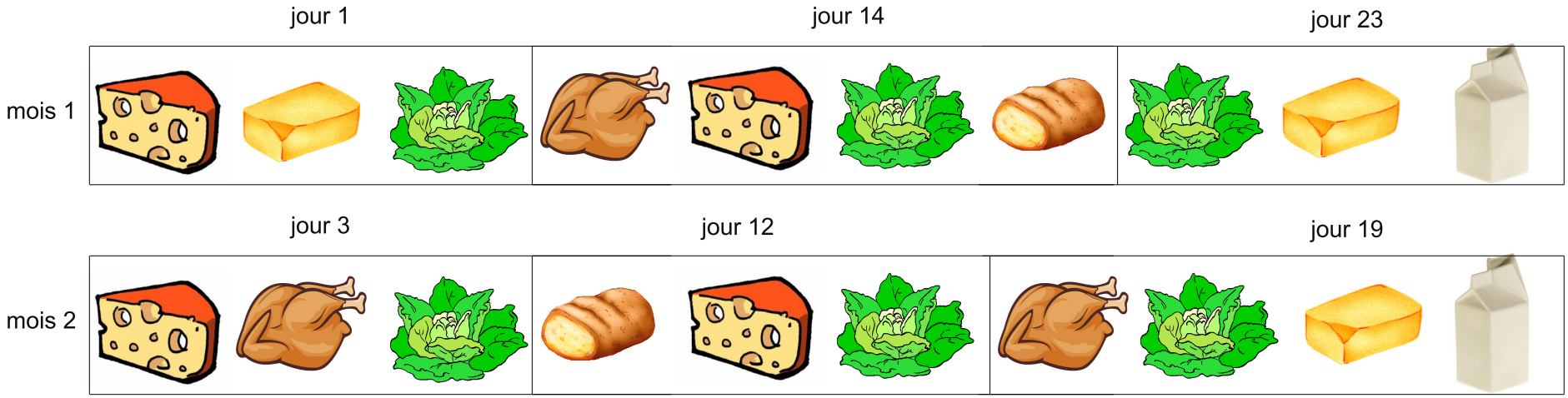


Achats fréquents (2 achats minimum) ?

- salade (3 achats),
- fromage suivi de salade (2 achats),
- beurre associé à salade n'est plus fréquent :
  - beurre suivi de salade (*1 seul achat*),
  - salade suivi de beurre (*1 seul achat*)

Exemple 2 : Prise en compte de la date de l'achat

Répartition des achats par mois



Définitions

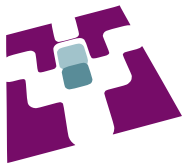
- Les produits sont appelés des items
- Les jours sont appelés des itemsets (**ensemble d'items**)
- Les achats par mois sont appelés des séquences (**liste ordonnée d'itemsets**)
- $\langle$*(fromage, salade)(fromage, salade, pain)(beurre, lait, salade)*$\rangle$: un motif séquentiel d'itemsets

Application : Analyse de textes biomédicaux

Modélisation en séquences d'items

Phrase : *"we conclude that* **VCP** *is essential for maturation of ubiquitin containing autophagosomes and that defect in this function may contribute to* **IBMPFD** *pathogenesis."*

Séquence
$\langle$(*we*#*PP*#) (*conclude*#*VBP*#) (*that*#*IN*#) (**GENE**) (*be*#*VBZ*#) (*essential*#*JJ*#) (*for*#*IN*#) (*maturation*#*NN*#) (*of*#*IN*#) (*ubiquitin*#*NN*#) (*contain*#*VBG*#) (*autophagosomes*#*NNS*#) (*and*#*CC*#) (*that*#*DT*#) (*defect*#*NN*#) (*in*#*IN*#) (*this*#*DT*#) (*function*#*NN*#) (*may*#*MD*#) (*contribute*#*VB*#) (*to*#*TO*#) (**DISEASE**) (*pathogenesis*#*NN*#)$\rangle$



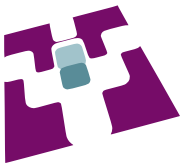
Rappel du sommaire

1. Contexte applicatif

2. Préliminaires
   - Les motifs ensemblistes
   - Les motifs séquentiels (MS)
   - L'extraction de MS fréquents

3. Contributions
   - Une contrainte globale pour l'extraction de MS fréquents
   - Une contrainte globale pour l'extraction de MS sous contrainte de gap

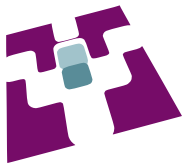
4. Perspectives



Extraction de motifs séquentiels : préliminaires

sid	Séquence
1	$\langle ABCBC \rangle$
2	$\langle BABC \rangle$
3	$\langle AB \rangle$
4	$\langle BCD \rangle$

- item : littéral, ex : $A, B, C, D, \ldots$
- séquence : liste ordonnée d'items, ex : $\langle ABCBC \rangle$
- base de séquences : ensemble de tuples (sid, s)
- sous-séquence : Un motif $s = \langle s_1 \ldots s_m \rangle$ est inclus dans un motif $s' = \langle s'_1 \ldots s'_n \rangle$ (noté) s'il existe des entiers $1 \leq j_1 < \ldots < j_m \leq n$ tels que $s_i = s'_{j_i}$, pour $1 \leq i \leq m$.
  ➥ $\langle AC \rangle$ est sous-séquence de $\langle ABCBC \rangle$
- support d'un motif
- extraction de motifs séquentiels



Extraction de motifs séquentiels (EMS)

sid	Séquence
1	$\langle ABCBC \rangle$
2	$\langle BABC \rangle$
3	$\langle AB \rangle$
4	$\langle BCD \rangle$

$\texttt{cover}(\langle AC \rangle) = \{(1, s_1), (2, s_2)\} \Rightarrow \texttt{sup}(\langle AC \rangle) = 2$

- séquence
- motif
- support d'un motif : nombres de séquences dans lequel apparaît le motif
- extraction de motifs séquentiels

Extraction de motifs séquentiels (EMS)

sid	Séquence
1	$\langle ABCBC \rangle$
2	$\langle BABC \rangle$
3	$\langle AB \rangle$
4	$\langle BCD \rangle$

avec $\theta = 2$ ($\sup(p) \geq 2$)

$FS = \{\langle A \rangle, \langle B \rangle, \langle C \rangle, \langle AB \rangle, \langle AC \rangle, \langle BB \rangle, \langle BC \rangle, \langle ABC \rangle, \langle \mathbf{BBC} \rangle\}$ ➠ 9 motifs !

- motif
- support d'un motif
- extraction de motifs fréquents : extraction de TOUS les motifs $>$ à un seuil (θ)

Extraction de motifs séquentiels (EMS)

sid	Séquence
1	$\langle ABCBC \rangle$
2	$\langle BABC \rangle$
3	$\langle AB \rangle$
4	$\langle BCD \rangle$

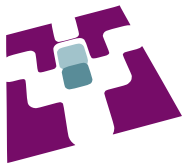
avec $\theta = 2$ ($\sup(p) \geq 2$)

$FS = \{\langle A \rangle, \langle B \rangle, \langle C \rangle, \langle AB \rangle, \langle AC \rangle, \langle BB \rangle, \langle BC \rangle, \langle ABC \rangle, \langle \mathbf{BBC} \rangle\}$ ➠ 9 motifs !

- motif
- support d'un motif
- extraction de motifs fréquents : extraction de TOUS les motifs $>$ à un seuil (θ)

Algorithmes d'extraction de MS

- Apriori : GSP (Srikant and Agrawal, EDBT'96)
- Pattern-growth : PrefixSpan (Pei et al., ICDE'01)
- Représentation verticale de la *SDB* : SPADE (Zaki, ML'01)
- EMS sous contraintes : SPIRIT (Garofalakis et al., VLDB'99) ; cSPADE (Zaki, CIKM'00) ; SMA (Trasarti et al., ICDM'08)
- MS fermés : CloSpan (Yan et al., SDM'03)



L'anti-monotonie de la contrainte de support minimal

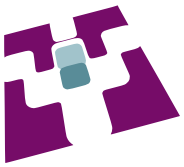
Propriété d'anti-monotonie (Agrawal & Sirkant, 94)

Une contrainte c est anti-monotone, si et seulement si, pour tout motif satisfaisant c, tous ses sous-motifs satisfont également c.

sid	Séquence
1	$\langle ABCBC \rangle$
2	$\langle BABC \rangle$
3	$\langle AB \rangle$
4	$\langle BCD \rangle$

Avec un seuil de support minimal $\theta = 2$

$\langle BA \rangle$ n'est pas fréquent $\rightarrow$ aucun sur-motif de $\langle BA \rangle$ ne pourra être fréquent



Fouille de motifs sous contraintes

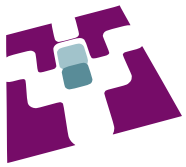
Pourquoi contraindre les motifs ?

- afin de **réduire** le nombre de motifs extraits
- cibler les motifs potentiellement intéressants

Approches existantes :

Contraintes	GSP	SPADE	cSPADE	SPAM	SPIRIT	BIDE	GAP-BIDE	PrefixSpan	TSP	TKS
Fréquence minimale	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Item										
Taille			✓						✓	✓
Gap			✓				✓			
Expressions régulières					✓					
Fermeture						✓	✓			
top-k									✓	✓

➠ Des méthodes dédiées selon les contraintes



EMS sous contraintes : contrainte de longueur minimale

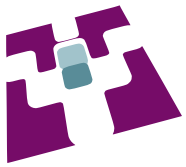
sid	Séquence
1	$\langle ABCBC \rangle$
2	$\langle BABC \rangle$
3	$\langle AB \rangle$
4	$\langle BCD \rangle$

Définition

$minSize(p, \ell_{min})$ impose que les motifs extraits p doivent contenir au moins ℓ_{min} items.

En imposant $\text{sup}(p) \geq 2 \wedge minSize(p, 3)$

seuls 2 motifs sont extraits : $\langle ABC \rangle$ and $\langle BBC \rangle$



EMS sous contraintes : contrainte d'item

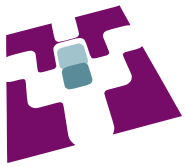
sid	Séquence
1	$\langle ABCBC \rangle$
2	$\langle BABC \rangle$
3	$\langle AB \rangle$
4	$\langle BCD \rangle$

Définition

$item(p, t)$ impose à ce que l'item t appartienne (ou pas) au motif p.

En imposant $\mathrm{sup}(p) \geq 3 \wedge maxSize(p, 2) \wedge item(p, C)$

seuls 2 motifs sont extraits : $\langle C \rangle$ and $\langle BC \rangle$



EMS sous contraintes : contrainte d'expression régulière

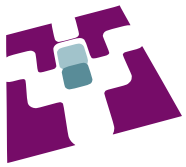
sid	Séquence
1	$\langle ABCBC \rangle$
2	$\langle BABC \rangle$
3	$\langle AB \rangle$
4	$\langle BCD \rangle$

Définition

$reg(p, exp)$ assure que le motif p doit être reconnu par un automate d'états finis associé à l'expression régulière exp.

En imposant $\text{sup}(p) \geq 2 \wedge reg(p, exp)$ où $exp = B\{BC|D\}$

le motif $\langle BBC \rangle$ est extrait de la base



EMS sous contraintes : top-k patterns

sid	Séquence
1	$\langle ABCBC \rangle$
2	$\langle BABC \rangle$
3	$\langle AB \rangle$
4	$\langle BCD \rangle$

Définition

Soient k et ℓ_{min} deux entiers strictement positifs. Un motif p est top-k motif de longueur minimale ℓ_{min} s'il existe au plus $(k-1)$ motifs ayant une longueur minimale ℓ_{min} et un support supérieur à celui de p.

avec $k = 5$ et $\ell_{min} = 1$

les *top*-5 motifs séquentiels : $\langle B \rangle : 4$, $\langle A \rangle : 3$, $\langle C \rangle : 3$, $\langle AB \rangle : 3$, et $\langle BC \rangle : 3$

La PPC pour la fouille séquentielle

Utiliser la PPC et ses solveurs génériques pour la fouille de motifs séquentiels sous contraintes :

- **Pro :** un cadre unifié et unique pour "composer" différentes contraintes
- **Cons :** problème du passage à l'échelle (en raison de l'encodage booléen utilisé)

  ➠ Un enjeu majeur pour les approches PPC existantes

Peu de travaux :

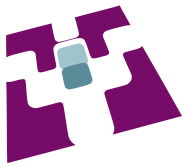
- Approche SAT [Coquery et al. (ECAI'12)]
- Modèles CSP [Métivier et al. (LML'13)], [Kemmar et al. (ICTAI'14)]
- Modèle PPC avec GC [Negrevergne et al. (CPAIOR'15)]

Principales limites de l'existant

Encodage de la base de séquences avec des contraintes réifiées
m variables booléennes S_s sont utilisées t.q.

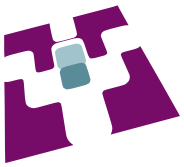
$$(S_s = 1) \Leftrightarrow (\text{p est sous-séquence de s})$$

- Pro :
  - Encodage de la contrainte de fréquence immédiat
    ➠ $sup_{SDB}(p) = \Sigma_{s \in SDB} S_s$
- Cons :
  - nécessite $(m = |SDB|)$ contraintes réifiées pour encoder toute la base
    ➠ Prohibitif même pour les bases de taille moyenne



Rappel du sommaire

1. Contexte applicatif
2. Préliminaires
   - Les motifs ensemblistes
   - Les motifs séquentiels (MS)
   - L'extraction de MS fréquents
3. Contributions
   - Une contrainte globale pour l'extraction de MS fréquents
   - Une contrainte globale pour l'extraction de MS sous contrainte de gap
4. Perspectives



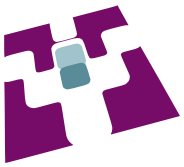
Une contrainte globale pour l'extraction de MS fréquents [Kemmar et al. (CP'15)]

Une nouvelle contrainte globale PREFIX-PROJECTION pour l'extraction de motifs séquentiels

➠ Peut être combinée avec d'autres contraintes

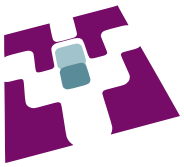
Contraintes	PP	SPADE	cSPADE	SPAM	SPIRIT	BIDE	GAP-BIDE	PrefixSpan
Fréquence minimale	✓	✓	✓	✓	✓	✓	✓	✓
Item	✓							
Taille	✓		✓					
Gap			✓				✓	
Expressions régulières	✓				✓			
Fermeture						✓	✓	
top-k	✓							

➠ Ne requiert ni contraintes ni variables supplémentaires pour encoder la relation de sous-séquence



Préfixe, Suffixe (Projection)

- La séquence $\alpha = \langle \alpha_1 \dots \alpha_m \rangle$ est un préfixe de $\beta = \langle \beta_1 \dots \beta_n \rangle$ $(m \leq n)$ ssi $\forall i \in [1..m], \alpha_i = \beta_i$
  - $\alpha = \langle AC \rangle$ est un préfixe de la séquence $\beta = \langle ACBC \rangle$
- Pour une séquence s, β est une projection de s par rapport à α ssi β est la plus grande sous-séquence de s ayant pour préfixe α
  - $\beta = \langle ACBC \rangle$ est la projection de $s = \langle ABCBC \rangle$ par rapport à α
- Séquence $\gamma = \langle \beta_{m+1} \dots \beta_n \rangle$ est appelée suffixe de s par rapport à α
  - $\langle BC \rangle$ est le suffixe de $\langle ABCBC \rangle$ par rapport à $\alpha = \langle AC \rangle$
- La α-projection, notée $SDB|_\alpha$, est l'ensemble de tous les suffixes des projections des séquences de SDB par rapport à α



PrefixSpan (Pei et al., 2001)

$\theta = 2$

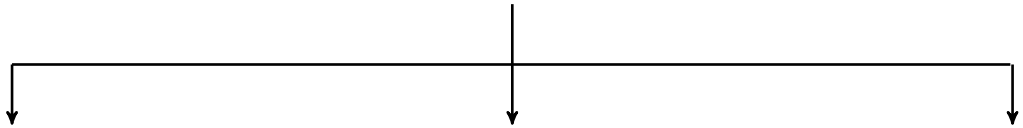
sid	Séquence
1	$\langle ABCBC \rangle$
2	$\langle BABC \rangle$
3	$\langle AB \rangle$
4	$\langle BCD \rangle$

length-1 seq. pat. :
$\langle A \rangle$, $\langle B \rangle$, $\langle C \rangle$

PrefixSpan (Pei et al., 2001)

$\theta = 2$

sid	Séquence
1	$\langle ABCBC \rangle$
2	$\langle BABC \rangle$
3	$\langle AB \rangle$
4	$\langle BCD \rangle$



length-1 seq. pat. :
$\langle A \rangle$, $\langle B \rangle$, $\langle C \rangle$

PrefixSpan (Pei et al., 2001)

$\theta = 2$

sid	Séquence
1	$\langle ABCBC \rangle$
2	$\langle BABC \rangle$
3	$\langle AB \rangle$
4	$\langle BCD \rangle$

length-1 seq. pat. :
$\langle A \rangle$, $\langle B \rangle$, $\langle C \rangle$

préfixe $\langle A \rangle$

$SDB\|_{\langle A \rangle}$
$\langle BCBC \rangle$
$\langle BC \rangle$
$\langle B \rangle$

préfixe $\langle B \rangle$

$SDB\|_{\langle B \rangle}$
$\langle CBC \rangle$
$\langle ABC \rangle$
$\langle \rangle$
$\langle CD \rangle$

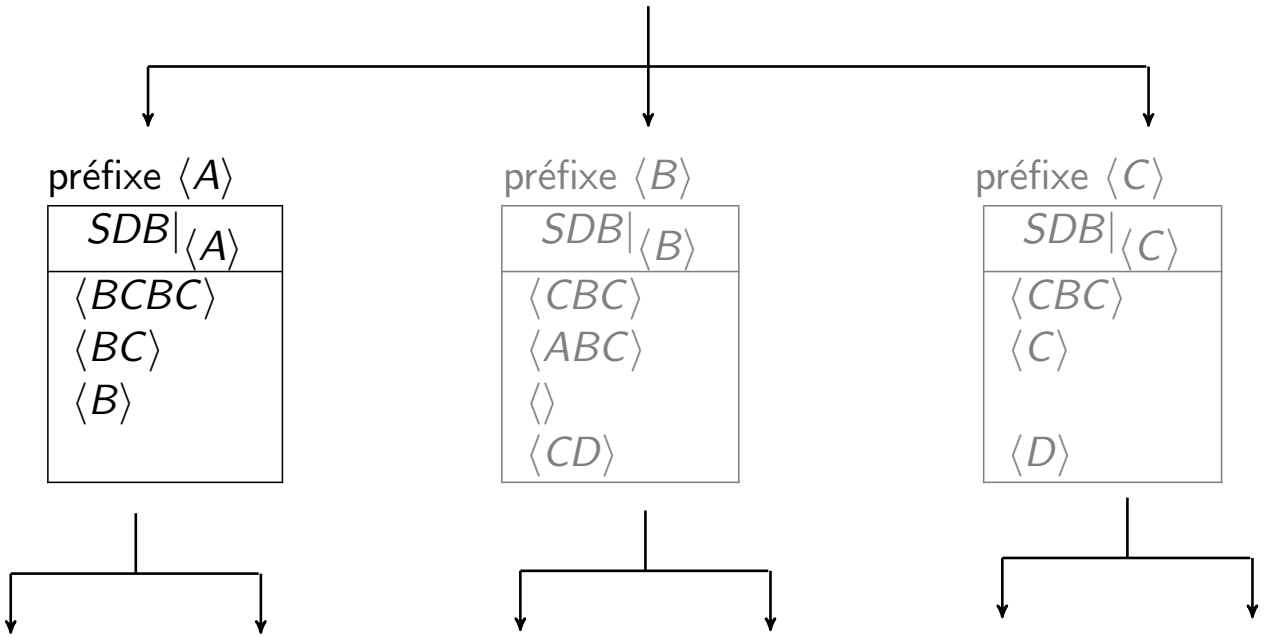
préfixe $\langle C \rangle$

$SDB\|_{\langle C \rangle}$
$\langle CBC \rangle$
$\langle C \rangle$
$\langle D \rangle$

PrefixSpan (Pei et al., 2001)

$\theta = 2$

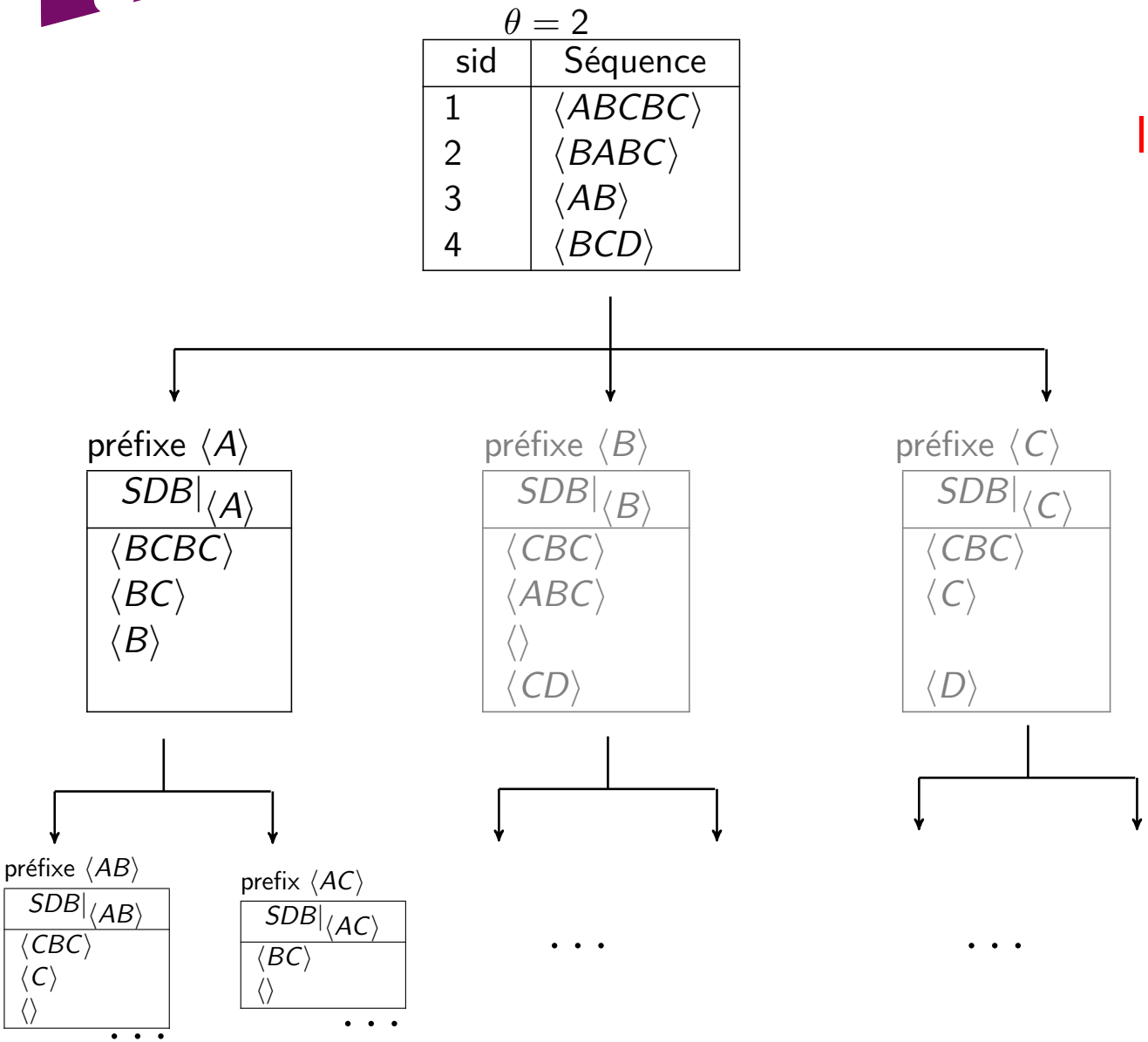
sid	Séquence
1	$\langle ABCBC \rangle$
2	$\langle BABC \rangle$
3	$\langle AB \rangle$
4	$\langle BCD \rangle$



length-1 seq. pat. :
$\langle A \rangle$, $\langle B \rangle$, $\langle C \rangle$

length-2 seq. pat. :
$\langle AB \rangle$, $\langle AC \rangle$, ...

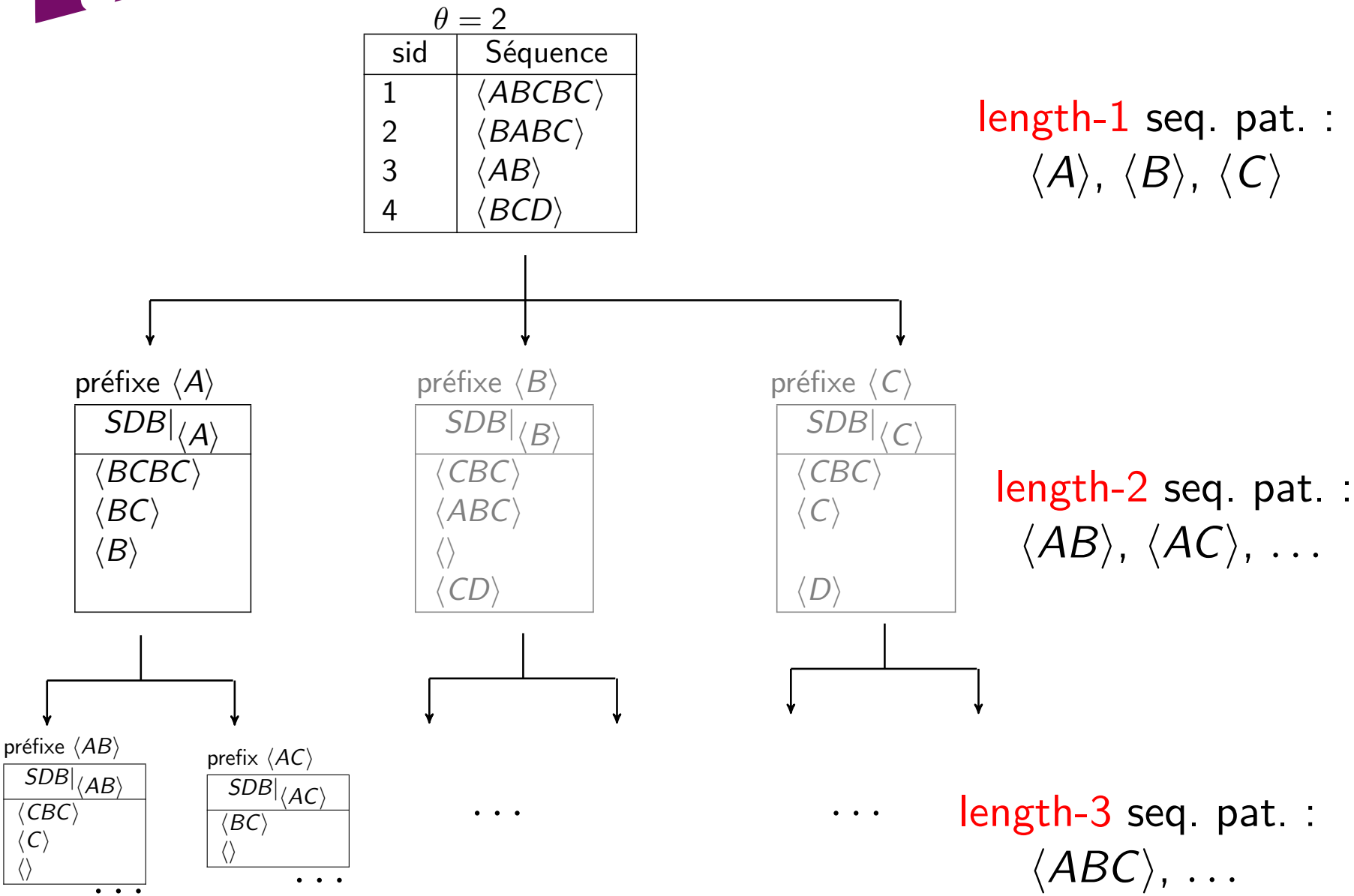
PrefixSpan (Pei et al., 2001)



length-1 seq. pat. :
$\langle A \rangle$, $\langle B \rangle$, $\langle C \rangle$

length-2 seq. pat. :
$\langle AB \rangle$, $\langle AC \rangle$, ...

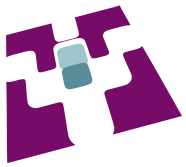
PrefixSpan (Pei et al., 2001)



length-1 seq. pat. :
$\langle A \rangle$, $\langle B \rangle$, $\langle C \rangle$

length-2 seq. pat. :
$\langle AB \rangle$, $\langle AC \rangle$, . . .

length-3 seq. pat. :
$\langle ABC \rangle$, . . .

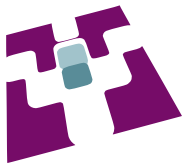


Calcul de support dans une base projetée

Calcul de support (Pei et al., 2001)

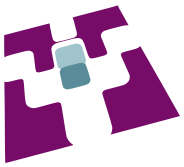
Pour toute séquence γ dans *SDB* ayant pour préfixe α et pour suffixe β, avec $\gamma = concat(\alpha, \beta)$, $sup_{SDB}(\gamma) = sup_{SDB|_\alpha}(\beta)$.

➠ Cette proposition garantit que seules les séquences dans *SDB* obtenues par extension du motif α seront considérées pour le calcul du support de la séquence γ. En outre, seuls les suffixes de $SDB|_\alpha$ doivent être considérés pour le calcul de ce support.



Prefix-Projection($[x_1, \ldots, x_\ell], \theta$) (1/3)

- $x = <x_1, \ldots, x_\ell>$ doit être un motif séquentiel t.q. $\sup(x) \geq \theta$
- Chaque variable x_i a pour domaine $D_i = \mathcal{I} \cup \{\square\}$
  - $\mathcal{I}$: ensemble de n items
  - $\square$: le symbole vide ($\square \notin \mathcal{I}$) indiquant la fin de la séquence
- Deux règles sur les domaines :
  - le premier item de x doit être non vide ($\square \notin D_1$)
  - $\forall i \in [2..(\ell - 1)], (x_i = \square) \rightarrow (x_{i+1} = \square)$



Prefix-Projection($[x_1, \ldots, x_\ell], \theta$) (2/3)

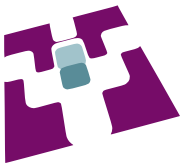
- Proposition 1 (Test de cohérence) :
  Solution de Prefix-Projection $([x_1, \ldots, x_\ell], \theta) \iff$
  affectation $\sigma = \langle d_1, ..., d_\ell \rangle$ des variables de x avec $\#SDB|_\sigma \geq \theta$

  ➠ $sup_{SDB}(\sigma) = sup_{SDB|_\sigma}(\langle\rangle) = |SDB|_\sigma|$

- Proposition 2 (Test de viabilité) :
  Soit σ une instanciation partielle consistante des variables $\langle x_1, \ldots, x_i \rangle$, et x_{i+1} une variable libre. Une valeur $d \in D(x_{i+1})$ participe à une solution de Prefix-Projection $([x_1, \ldots, x_\ell], \theta)$ ssi d est un item fréquent dans $SDB|_\sigma$: $|\{(sid, \gamma) | (sid, \gamma) \in SDB|_\sigma \wedge \langle d \rangle \preceq \gamma\}| \geq \theta$

- Filtrage (principe/idée) :
  S'il n'existe aucun suffixe s débutant par it t.q. $\text{sup}(s) \geq \theta$ alors l'item it peut être retiré des domaines $D_i, \ldots, D_\ell$

- mise en œuvre : projection préfixée [Pei et al. (ICDE'01)]



Prefix-Projection($[x_1, \ldots, x_n], \theta$) (3/3)

Exemple :

- ensemble d'items $\mathcal{I} = \{A, B, C, D\}$, $\theta = 2$ et $x = < x_1, x_2, x_3 >$

sid	Séquence
1	$ABCDBC$
2	$BABC$
3	AB
4	BCD

- soit $x_1 = A$
- projection préfixée selon A :
  $SDB|_A = \{(1, \langle BCDBC \rangle), (2, \langle BC \rangle), (3, \langle B \rangle)\}$
- seuls les suffixes s débutant par B ou C pourront satisfaire $\text{sup}(s) \geq 2$
- donc $x_2 \neq A$, $x_2 \neq D$ et idem pour x_3

  ➠ $D(x_2) = D(x_3) = \{B, C, \square\}$

Algorithme de filtrage (1/2)

Algorithme 1 : Filter-Prefix-Projection(SDB, σ, i, x, $minsup$)

begin

1 **if** $(i \geq 2 \wedge \sigma(x_i) = \square)$ **then**

2 **for** $j \leftarrow i+1$ **to** ℓ **do**

3 $P_j \leftarrow \square$;

4 **return** True;

else

5 $\mathcal{PSDB}_i \leftarrow \texttt{ProjectSDB}(SDB, \mathcal{PSDB}_{i-1}, \langle\sigma(x_i)\rangle)$;

6 **if** $(\#\mathcal{PSDB}_i < minsup)$ **then**

7 **return** False ;

else

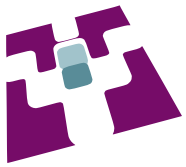
8 $\mathcal{FI} \leftarrow \texttt{getFreqItems}(SDB, \mathcal{PSDB}_i, minsup)$;

9 **for** $j \leftarrow i+1$ **to** ℓ **do**

10 **foreach** $a \in D(x_j)\, s.t. (a \neq \square \wedge a \notin \mathcal{FI})$ **do**

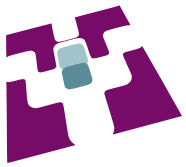
11 $D(x_j) \leftarrow D(x_j) - \{a\}$;

12 **return** True;



Algorithme de filtrage (2/2)

- Filtrage
  - cohérence de domaine sur la variable x_{i+1} qui suit le préfixe courant
  - structures de données incrémentales pour calculer les projections préfixées intermédiaires
- Complexités
  - en temps : $O(m \times \ell + m \times n + \ell \times n)$
  - en espace : $O(m \times \ell)$
- Recherche de solutions
  - L'extraction de TOUS les motifs séquentiels de cardinalité K, se fait **sans échec**, et en $O(K \times \ell \times (m \times \ell + m \times n + \ell \times n))$



Encodage des contraintes locales sur les motifs

1. Contrainte d'items : $item(x, \mathcal{V}) \equiv \bigwedge_{t \in \mathcal{V}} \texttt{Among}(x, \{t\}, l, u)$
   - ➠ les items de $\mathcal{V}$ doivent apparaître au moins l fois et au plus u fois dans x
2. Contrainte de longueur minimale : $minSize(x, \ell_{min}) \equiv \bigwedge_{i=1}^{i=\ell_{min}} (x_i \neq \square)$
3. Contrainte de longueur maximale : $maxSize(x, \ell_{max}) \equiv \bigwedge_{i=\ell_{max}+1}^{i=\ell} (x_i = \square)$
4. Contrainte d'expression régulière : $reg(x, exp) \equiv \texttt{Regular}(x, A_{reg})$

Résultats expérimentaux

- Datasets :

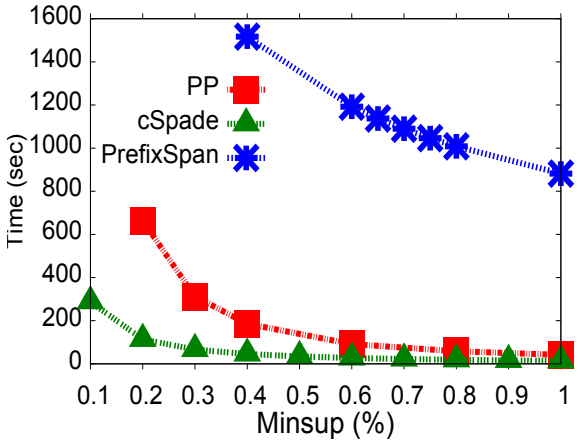
dataset	$\|SDB\|$	$\|\mathcal{I}\|$	avg ($\|s\|$)	$\max_{s \in SDB}$ ($\|s\|$)	type de données
Leviathen	5834	9025	33.81	100	book
FIFA	20450	2990	34.74	100	web click stream
BIBLE	36369	13905	21.64	100	bible
Kosarak	69999	21144	7.97	796	web click stream
Protein	103120	24	482	600	protein sequences
PubMed	17527	19931	29	198	bio-medical text
data-200K	200000	20	50	86	synthetic dataset

- Implémentation : `Gecode`, timeout = 3600 `secondes`
  ➥ `https://sites.google.com/site/prefixprojection4cp/`

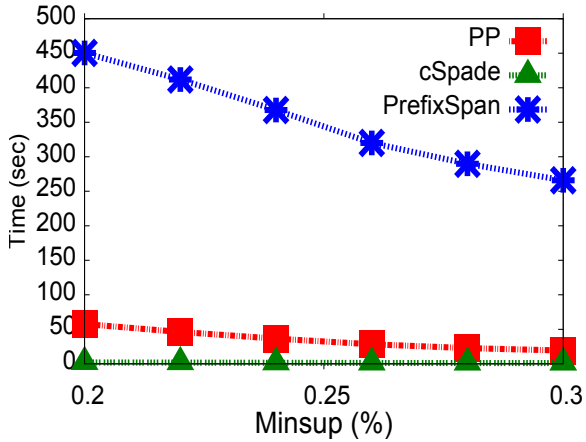
Résultats expérimentaux

Comparaison avec les méthodes spécialisées

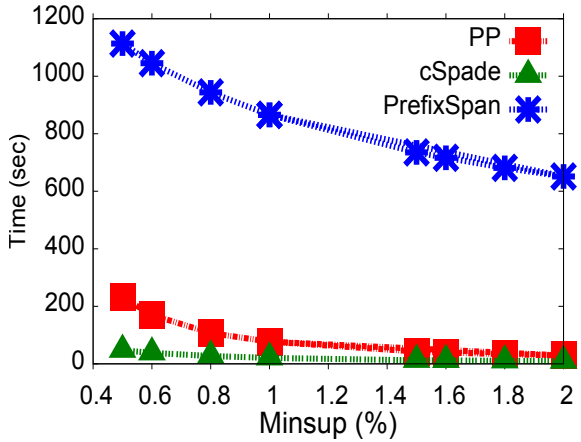
BIBLE



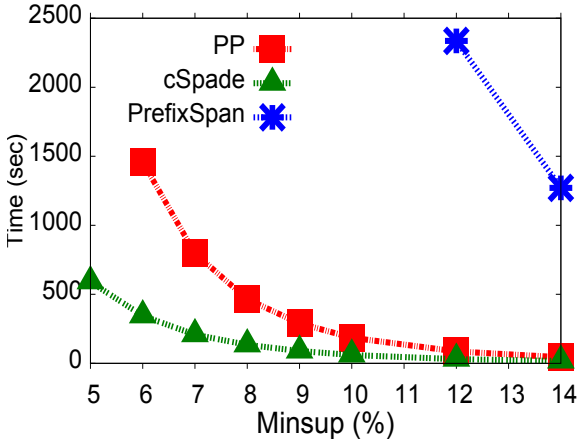
Kosarak



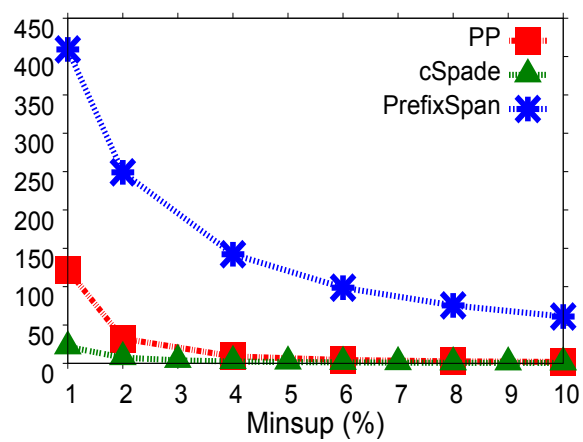
PubMed



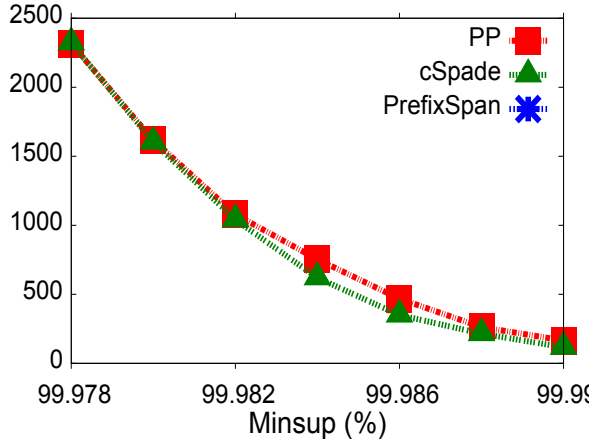
FIFA



Leviathan



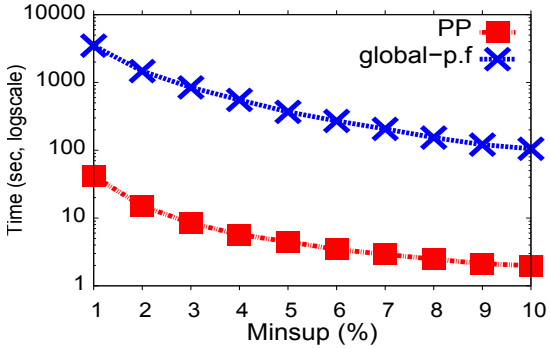
Protein



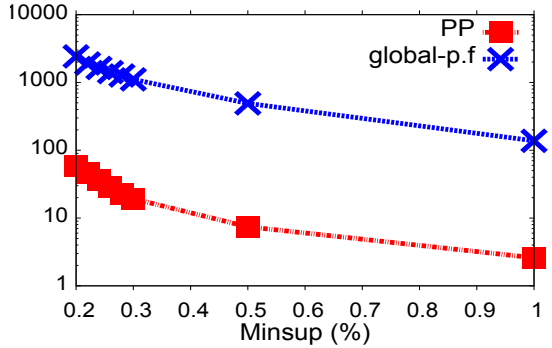
Résultats expérimentaux

Comparaison avec les méthodes CP

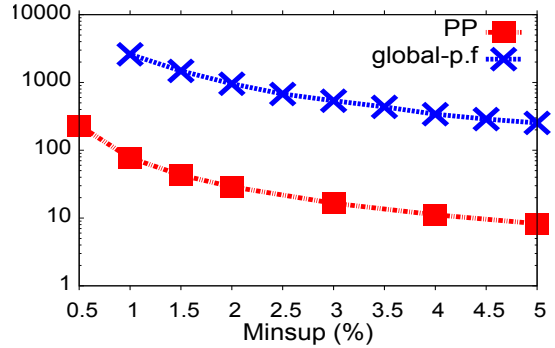
BIBLE



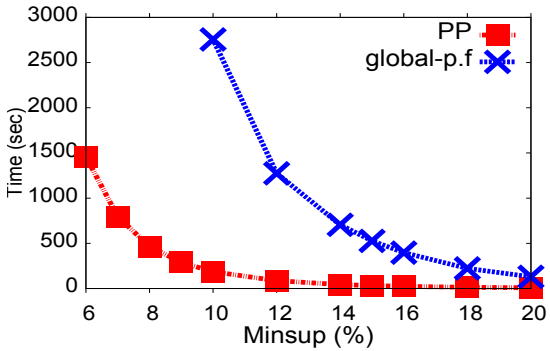
Kosarak



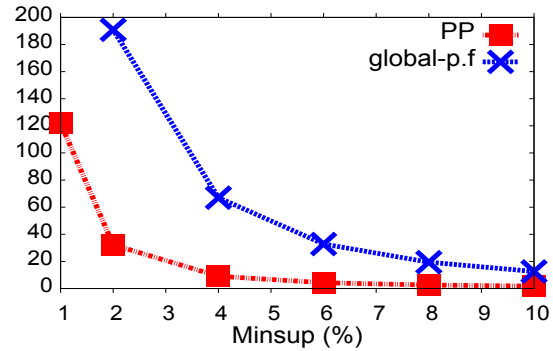
PubMed



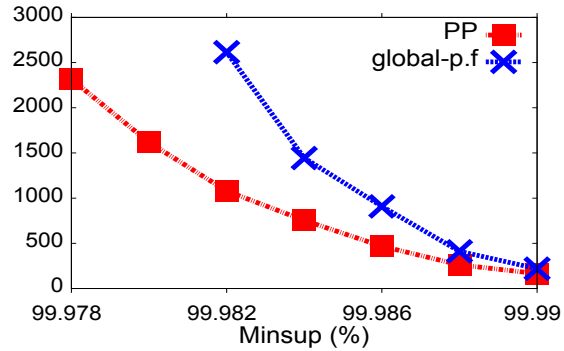
FIFA



Leviathan



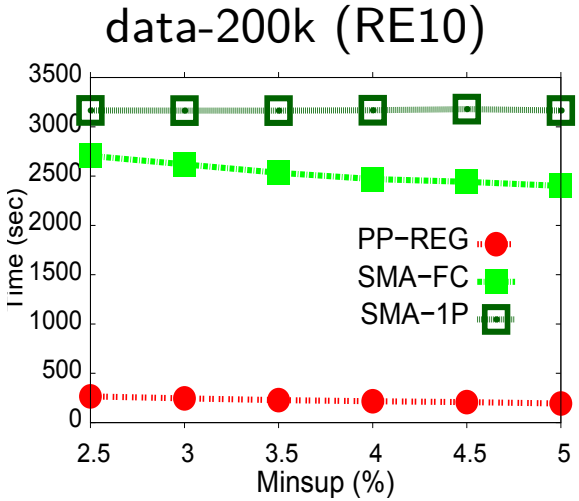
Protein

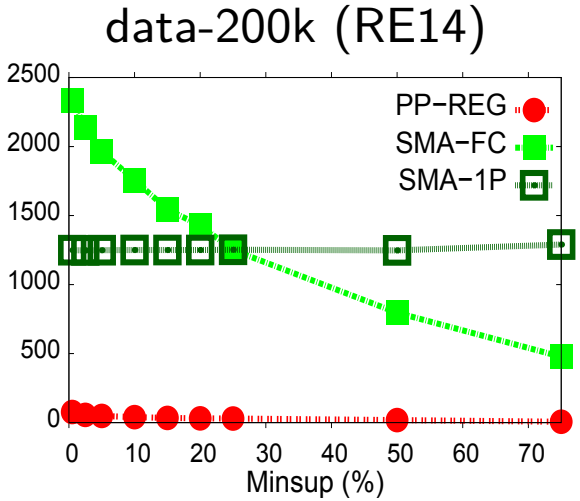


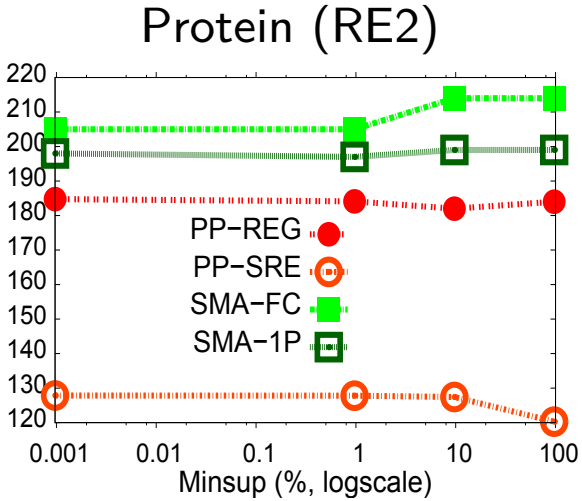
- `PP` is more than an order of magnitude faster than `global-p.f`
- `global-p.f` is not able to mine for patterns at very low frequency within the time limit

Résultats expérimentaux

Prefix-Projection avec contraintes d'ER

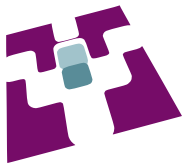






- PP−REG est au moins un ordre de grandeur plus rapide que SMA [Bonchi et al. (ICDM 2008)]

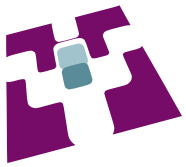
- RE10 $\equiv A^*B(B|C)D^*EF^*(G|H)I^*$
- RE14 $\equiv A^*(Q|BS^*(B|C))D^*E(I|S)^*(F|H)G^*R$
- RE2 $\equiv (S|T)\ .\ (R|K)$



Autres résultats

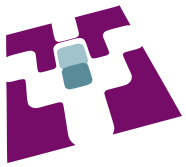
top-k motifs séquentiels [Kemmar et al. (Constraints'16)]

- exploitation de PREFIX-PROJECTION pour explorer l'espace des motifs séquentiels
- **Extraction en deux étapes principales** :
  - étape d'initialisation : extraction des k premiers motifs
  - étape de "support raising" : mise à jour du seuil minimum θ durant la recherche ➠ au départ, $\theta = 0$
- Proposition d'une **nouvelle heuristique** pour booster la phase d'initialisation ➠ augmenter θ plus rapidement
- **Temps de calcul :** notre approche domine largement la méthode de l'état de l'art TSP [Tzvetkov et al. 'ICDM'03)] et surpasse TKS [Fournier-Viger et al. 2013] sur certains jeux de données.



Rappel du sommaire

1. Contexte applicatif
2. Préliminaires
   - Les motifs ensemblistes
   - Les motifs séquentiels (MS)
   - L'extraction de MS fréquents
3. Contributions
   - Une contrainte globale pour l'extraction de MS fréquents
   - Une contrainte globale pour l'extraction de MS sous contrainte de gap
4. Perspectives



EMS sous contraintes : contrainte de gap

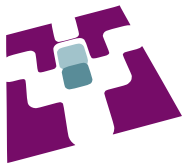
sid	Séquence
1	$\langle ABCDB \rangle$
2	$\langle ACCBACB \rangle$
3	$\langle ADCBEEC \rangle$
4	$\langle AACC \rangle$

Définition

Un motif α satisfaisant la contrainte $gap[M, N]$ est un motif tel que au moins M items et au plus N items sont autorisés entre chaque paire d'items adjacents, dans les séquences d'origines s (noté $\alpha \preceq^{[M,N]} s$).

Exemple avec $gap[0, 1]$

$\langle AC \rangle \preceq^{[0,1]} \langle ABCDB \rangle$. La paire $(s_1, [1, 3])$ dénote l'**occurrence** de $\langle AC \rangle$ dans $\langle ABCDB \rangle$.

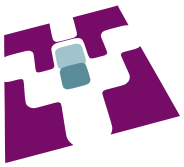


EMS sous contraintes : contrainte de gap

- $AllOcc(\alpha, s)$: l'ensemble de toutes les occurrences d'une séquence α sous $gap[M, N]$ dans la séquence s.
- $AllOcc(\alpha, SDB)$: l'ensemble de toutes les occurrences d'une séquence α sous $gap[M, N]$ dans la base de séquences SDB.
- $cover_{SDB}^{[M,N]}(\alpha)$: l'ensembles de toutes les séquences de SDB dans lesquels α est contenu.
- $sup_{SDB}^{[M,N]}(\alpha)$: le support de la séquence α dans SDB.

Exemple avec $gap[0, 1]$ et $\alpha = \langle AC \rangle$

$AllOcc(\alpha, \langle ACCBACB \rangle) = \{[1, 2]), [1, 3], [5, 6]\}$.
$cover_{SDB_1}^{[0,1]}(\alpha) = \{(1, s_1), (2, s_2), (3, s_3), (4, s_4)\}$.

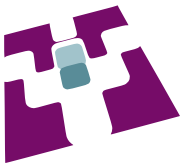


Propriété de préfixe anti-monotonie de $gap[M, N]$

La propriété d'antimonotonie du support n'est plus vérifiée avec gap :

sid	Sequence
1	$\langle ABCDB \rangle$
2	$\langle ACCBACB \rangle$
3	$\langle ADCBEEC \rangle$
4	$\langle AACC \rangle$

- Le motif $\langle AB \rangle$ n'est pas fréquent sous la contrainte $gap[0, 1]$ (pour $\theta = 3$) alors que le motif $\langle ACB \rangle$ est un motif fréquent sous la contrainte $gap[0, 1]$.



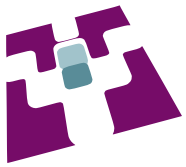
Propriété de préfixe anti-monotonie de $gap[M, N]$

Propriété de préfixe anti-monotonie (Pei et al., 2002)

Une contrainte c est dite préfixe anti-monotone, si et seulement si, pour tout motif α vérifiant c, chaque préfixe de α satisfait également c.

Proposition 1 : $gap[M, N]$ est préfixe anti-monotone.

➠ si une séquence α ne satisfait pas la contrainte de $gap[M, N]$, alors toutes les séquences qui possèdent α comme préfixe ne peuvent pas satisfaire cette contrainte.



Right pattern extensions (1/2)

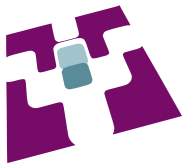
Definition

Soit une séquence (sid, s) et un motif p t.q. $p \preceq^{[M,N]} s$. Les **extensions droites** de p dans s, notées $Ext_R^{[M,N]}(p, s)$, est l'ensemble des sous-séquences légales de s qui permettent d'étendre à droite p afin de former un motif satisfaisant la contrainte $gap[M, N]$.

➠ $Ext_R^{[M,N]}(p, s)$: L'union de toutes les sous-séquences de s qui sont dans l'intervalle $[j_m + M + 1, min(j_m + N + 1, \#s)]$, avec $[j_1, j_m] \in AllOcc(p, s)$

Exemple avec $gap[0, 1]$ et $\alpha = \langle AC \rangle$

$Ext_R^{[0,1]}(\langle AC \rangle, \langle ACCBACB \rangle) = \{\langle CB \rangle, \langle BA \rangle, \langle B \rangle\}$



Right pattern extensions (1/2)

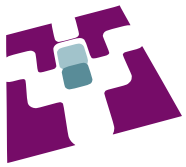
Definition

Soit une séquence (sid, s) et un motif p t.q. $p \preceq^{[M,N]} s$. Les **extensions droites** de p dans s, notées $Ext_R^{[M,N]}(p, s)$, est l'ensemble des sous-séquences légales de s qui permettent d'étendre à droite p afin de former un motif satisfaisant la contrainte $gap[M, N]$.

➠ $Ext_R^{[M,N]}(p, s)$: L'union de toutes les sous-séquences de s qui sont dans l'intervalle $[j_m + M + 1, min(j_m + N + 1, \#s)]$, avec $[j_1, j_m] \in AllOcc(p, s)$

Exemple avec $gap[0, 1]$ et $\alpha = \langle AC \rangle$

$Ext_R^{[0,1]}(\langle AC \rangle, \langle ACCBACB \rangle) = \{\langle CB \rangle, \langle BA \rangle, \langle B \rangle\}$



Right pattern extensions (1/2)

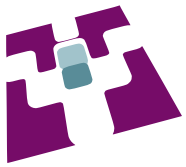
Definition

Soit une séquence (sid, s) et un motif p t.q. $p \preceq^{[M,N]} s$. Les **extensions droites** de p dans s, notées $Ext_R^{[M,N]}(p, s)$, est l'ensemble des sous-séquences légales de s qui permettent d'étendre à droite p afin de former un motif satisfaisant la contrainte $gap[M, N]$.

➠ $Ext_R^{[M,N]}(p, s)$: L'union de toutes les sous-séquences de s qui sont dans l'intervalle $[j_m + M + 1, min(j_m + N + 1, \#s)]$, avec $[j_1, j_m] \in AllOcc(p, s)$

Exemple avec $gap[0, 1]$ et $\alpha = \langle AC \rangle$

$Ext_R^{[0,1]}(\langle AC \rangle, \langle ACCBACB \rangle) = \{\langle CB \rangle, \langle BA \rangle, \langle B \rangle\}$



Right pattern extensions (1/2)

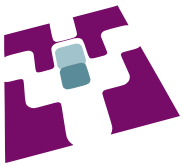
Definition

Soit une séquence (sid, s) et un motif p t.q. $p \preceq^{[M,N]} s$. Les **extensions droites** de p dans s, notées $Ext_R^{[M,N]}(p, s)$, est l'ensemble des sous-séquences légales de s qui permettent d'étendre à droite p afin de former un motif satisfaisant la contrainte $gap[M, N]$.

➠ $Ext_R^{[M,N]}(p, s)$: L'union de toutes les sous-séquences de s qui sont dans l'intervalle $[j_m + M + 1, min(j_m + N + 1, \#s)]$, avec $[j_1, j_m] \in AllOcc(p, s)$

Exemple avec $gap[0, 1]$ et $\alpha = \langle AC \rangle$

$Ext_R^{[0,1]}(\langle AC \rangle, \langle ACCBACB \rangle) = \{\langle CB \rangle, \langle BA \rangle, \langle B \rangle\}$



Right pattern extensions (2/2)

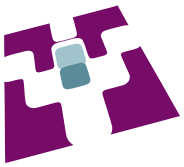
Les **extensions droites de p dans la base de séquences SDB** :

$$Ext_R^{[M,N]}(p, SDB) = \bigcup_{(sid,s)\in SDB} \{(sid, Ext_R^{[M,N]}(p, s))\}$$

➠ Cas avec la contrainte gap :

sid	Sequence
1	$\langle ABCDB\rangle$
2	$\langle ACCBACB\rangle$
3	$\langle ADCBEEC\rangle$
4	$\langle AACC\rangle$

$$Ext_R^{[0,1]}(\langle AC\rangle, SDB_1) = \{(1, \{\langle DB\rangle\}), (2, \{\langle CB\rangle, \langle BA\rangle, \langle B\rangle\}), (3, \{\langle BE\rangle\}), (4, \{\langle C\rangle\})\}$$



Right pattern extensions (2/2)

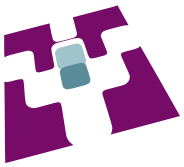
Les **extensions droites de p dans la base de séquences SDB** :

$$Ext_R^{[M,N]}(p, SDB) = \bigcup_{(sid,s)\in SDB} \{(sid, Ext_R^{[M,N]}(p, s))\}$$

➠ Cas sans contrainte gap ($N \geq \#s$) :
comme toute extension de p atteint toujours la fin de la séquence s, considérer uniquement l'extension ayant la **plus petite position valide**

sid	Sequence
1	$\langle ABCDB \rangle$
2	$\langle ACCBACB \rangle$
3	$\langle ADCBEEC \rangle$
4	$\langle AACC \rangle$

$$Ext_R^{[0,\infty]}(\langle AC \rangle, SDB_1) = \{(1, \{\langle DB \rangle\}), (2, \{\langle CBACB \rangle\}), (3, \{\langle BEEC \rangle]\}), (4, \{\langle C \rangle\})\}$$



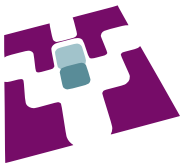
Calcul du support sous $gap[M, N]$

Proposition 2 : Un motif p est un motif fréquent sous la contrainte $gap[M, N]$ dans la base de séquences SDB, si et seulement si, la condition suivante est vérifiée : $\#Ext_R^{[M,N]}(p, SDB) \geq \theta$

$\#Ext_R^{[0,1]}(\langle AC \rangle, SDB_1) = 4 > minsup = 2$

➠ $\langle AC \rangle$ est un motif fréquent sous la contrainte $gap[0, 1]$.

Résultat principal : L'ensemble des items localement fréquents $\mathcal{RF}^{[M,N]}(p, SDB)$ dans les extensions droites de p dans SDB peut être exploité pour étendre p.



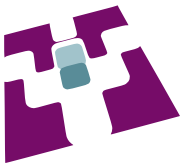
GAP-SEQ $([x_1, \ldots, x_\ell], \theta, M, N)$ (1/2)

- $x = <x_1, \ldots, x_\ell>$ doit être un motif séquentiel t.q. $sup_{SDB}^{[M,N]}(x) \geq \theta$

- Chaque variable x_i a pour domaine $D_i = \mathcal{I} \cup \{\square\}$
  - $\mathcal{I}$: ensemble de n items
  - $\square$: le symbole vide ($\square \notin \mathcal{I}$) indiquant la fin de la séquence

- Proposition 3 (Test de cohérence) :
  Solution de GAP-SEQ $([x_1, \ldots, x_\ell], \theta, M, N) \iff$
  affectation $\sigma = \langle d_1, ..., d_\ell \rangle$ des variables de x t.q. $\#Ext_R^{[M,N]}(\sigma, SDB) \geq \theta$

- Proposition 4 : Soit σ une instanciation partielle consistante des variables $\langle x_1, \ldots, x_i \rangle$, et x_{i+1} une variable libre. Une valeur $d \in D(x_{i+1})$ participe à une solution de GAP-SEQ $([x_1, \ldots, x_\ell], \theta, M, N)$ ssi $d \in \mathcal{RF}^{[M,N]}(\sigma, SDB)$.



GAP-SEQ $([x_1, \ldots, x_\ell], \theta, M, N)$ (2/2)

Exemple :

- ensemble d'items $\mathcal{I} = \{A, B, C, D, E\}$, $\theta = 2$ et $gap[1, 2]$

sid	Sequence
1	$\langle ABCDB \rangle$
2	$\langle ACCBACB \rangle$
3	$\langle ADCBEEC \rangle$
4	$\langle AACC \rangle$

- $x = < x_1, x_2, x_3, x_4 >$ avec $D(x_1) = \mathcal{I}$, $D(x_2) = D(x_3) = D(x_4) = \mathcal{I} \cup \{\square\}$
- $x_1 = A$

GAP-SEQ $([x_1, \ldots, x_\ell], \theta, M, N)$ (2/2)

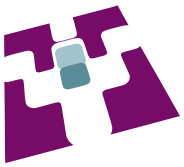
Exemple :

- ensemble d'items $\mathcal{I} = \{A, B, C, D, E\}$, $\theta = 2$ et $gap[1, 2]$

sid	Sequence
1	$\langle ABCDB \rangle$
2	$\langle ACCBACB \rangle$
3	$\langle ADCBEEC \rangle$
4	$\langle AACC \rangle$

- $x = < x_1, x_2, x_3, x_4 >$ avec $D(x_1) = \mathcal{I}$, $D(x_2) = D(x_3) = D(x_4) = \mathcal{I} \cup \{\square\}$
- $x_1 = A$

- $Ext_R^{[1,2]}(\langle A \rangle, SDB_1) = \{(1, \{\langle CD \rangle\}), (2, \{\langle CB \rangle, \langle B \rangle\}), (3, \{\langle CB \rangle\}), (4, \{\langle CC \rangle, \langle C \rangle\})\}$



GAP-SEQ $([x_1, \ldots, x_\ell], \theta, M, N)$ (2/2)

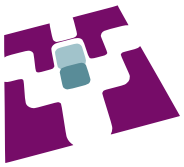
Exemple :

- ensemble d'items $\mathcal{I} = \{A, B, C, D, E\}$, $\theta = 2$ et $gap[1, 2]$

sid	Sequence
1	$\langle ABCDB \rangle$
2	$\langle ACCBACB \rangle$
3	$\langle ADCBEEC \rangle$
4	$\langle AACC \rangle$

- $x = < x_1, x_2, x_3, x_4 >$ avec $D(x_1) = \mathcal{I}$, $D(x_2) = D(x_3) = D(x_4) = \mathcal{I} \cup \{\square\}$
- $x_1 = A$

- $Ext_R^{[1,2]}(\langle A \rangle, SDB_1) = \{(1, \{\langle CD \rangle\}), (2, \{\langle CB \rangle, \langle B \rangle\}), (3, \{\langle CB \rangle\}), (4, \{\langle CC \rangle, \langle C \rangle\})\}$
- Items localement fréquent dans $Ext_R^{[1,2]}(\langle A \rangle, SDB_1)$: $\{B, C\}$



GAP-SEQ $([x_1, \ldots, x_\ell], \theta, M, N)$ (2/2)

Exemple :

- ensemble d'items $\mathcal{I} = \{A, B, C, D, E\}$, $\theta = 2$ et $gap[1, 2]$

sid	Sequence
1	$\langle ABCDB \rangle$
2	$\langle ACCBACB \rangle$
3	$\langle ADCBEEC \rangle$
4	$\langle AACC \rangle$

- $x = < x_1, x_2, x_3, x_4 >$ avec $D(x_1) = \mathcal{I}$, $D(x_2) = D(x_3) = D(x_4) = \mathcal{I} \cup \{\square\}$
- $x_1 = A$

- $Ext_R^{[1,2]}(\langle A \rangle, SDB_1) = \{(1, \{\langle CD \rangle\}), (2, \{\langle CB \rangle, \langle B \rangle\}), (3, \{\langle CB \rangle\}), (4, \{\langle CC \rangle, \langle C \rangle\})\}$
- Items localement fréquent dans $Ext_R^{[1,2]}(\langle A \rangle, SDB_1)$: $\{B, C\}$
- Supprimer les valeurs A, D et E de $D(x_2)$
  - ➠ $D(x_2) = \{B, C, \square\}$

Algorithme de filtrage

Algorithme 2 : FILTER-GAP-SEQ(SDB, σ, j, x, $minsup$, M, N)

begin

  $Ext_R \leftarrow$ `getRightExt`$(SDB, \mathcal{ALLOCC}_{j-1}, \sigma, M, N)$;

  if $(\#Ext_R < minsup)$ **then**

    return False ;

  if $(j \geq 2 \wedge \sigma(x_j) = \square)$ **then**

    for $k \leftarrow j+1$ **to** ℓ **do**

      $x_k \leftarrow \square$;

  else

    $\mathcal{RF} \leftarrow$ `getFreqItems`$(SDB, Ext_R, minsup)$;

    foreach $a \in D(x_{j+1})\, s.t. (a \neq \square \wedge a \notin \mathcal{RF})$ **do**

      $D(x_{j+1}) \leftarrow D(x_{j+1}) - \{a\}$;

  return True ;

Complexités :

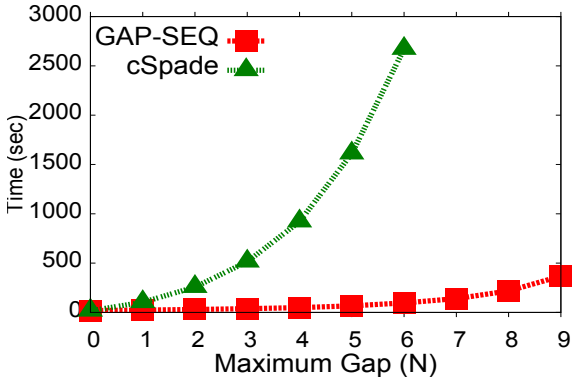
- en temps : $O(m \times \ell^2 + d)$
- en espace : $O(m \times \ell^2)$

Résultats expérimentaux (1/5)

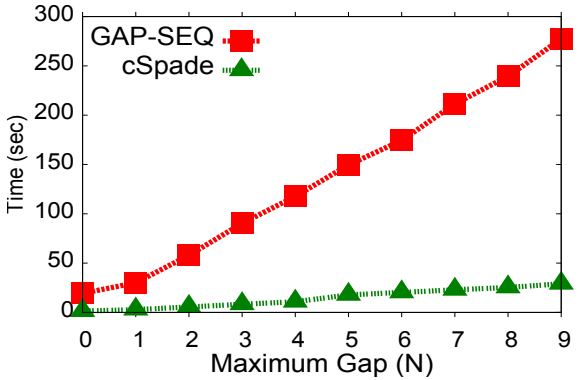
Comparaison avec cSpade : runtime

Varying the value of parameter N in the gap constraint ($M = 0$)

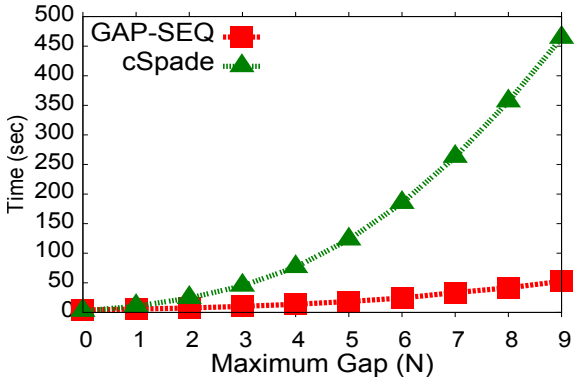
BIBLE (0.1%)



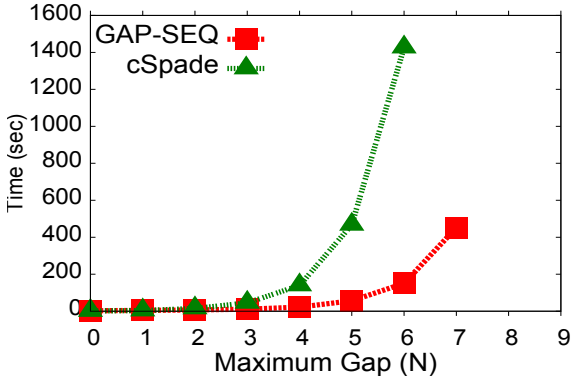
Kosarak (0.1%)



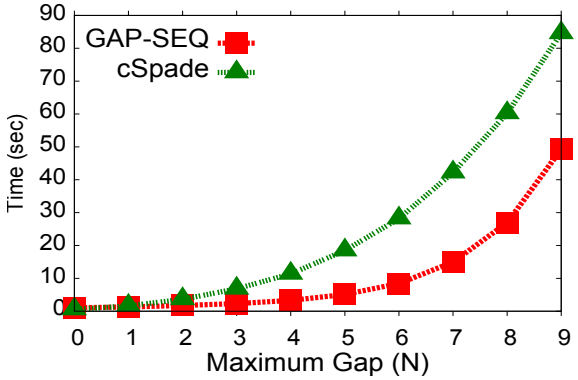
PubMed (0.5%)



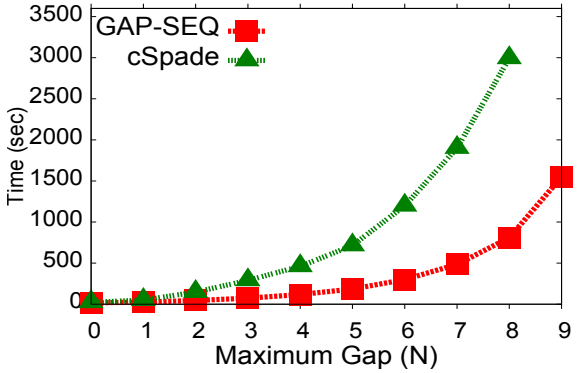
FIFA (2%)



Leviathan (0.8%)



Protein (96%)

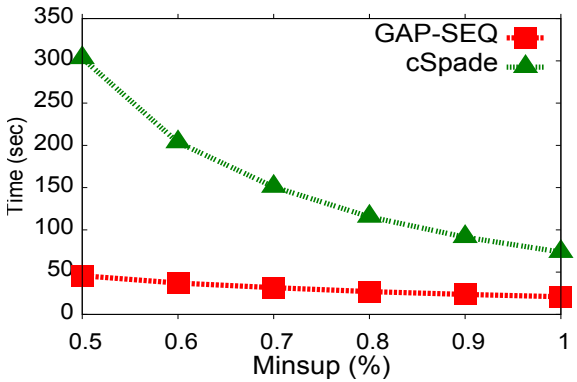


Résultats expérimentaux (2/5)

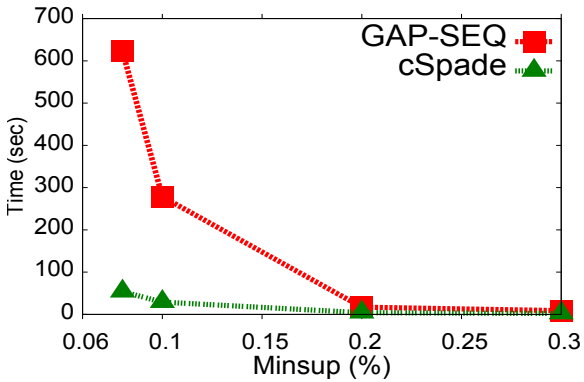
Comparaison avec cSpade : runtime

Varying the value of θ with the gap constraint $gap[0, 9]$

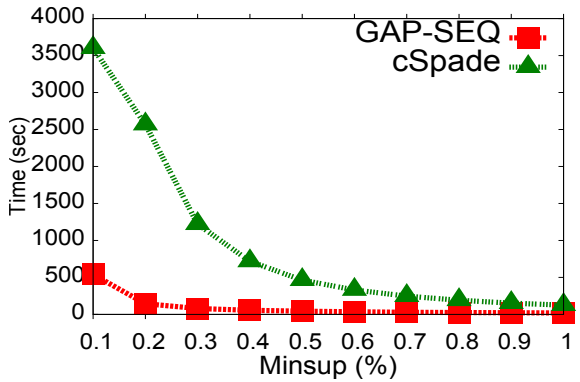
BIBLE



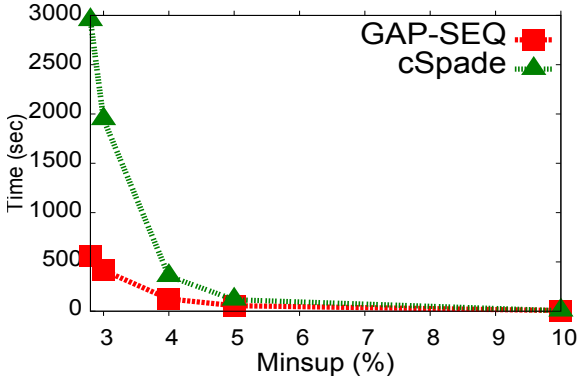
Kosarak



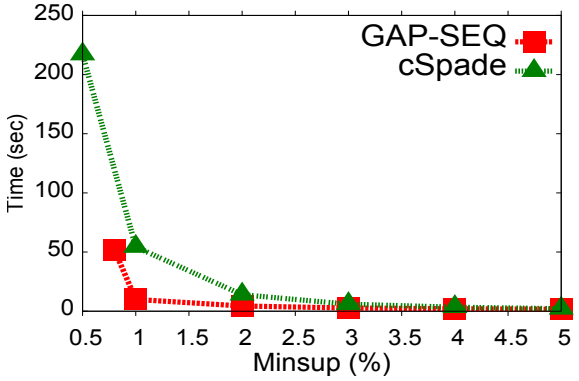
PubMed



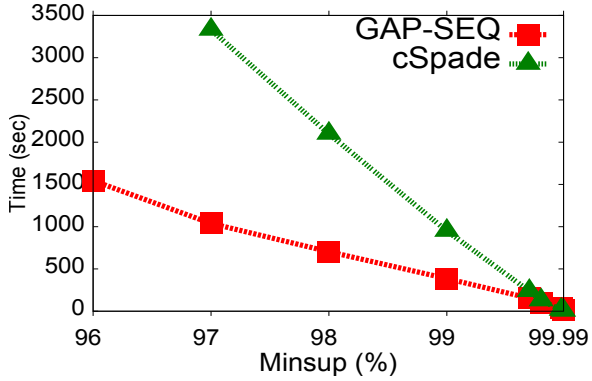
FIFA



Leviathan



Protein

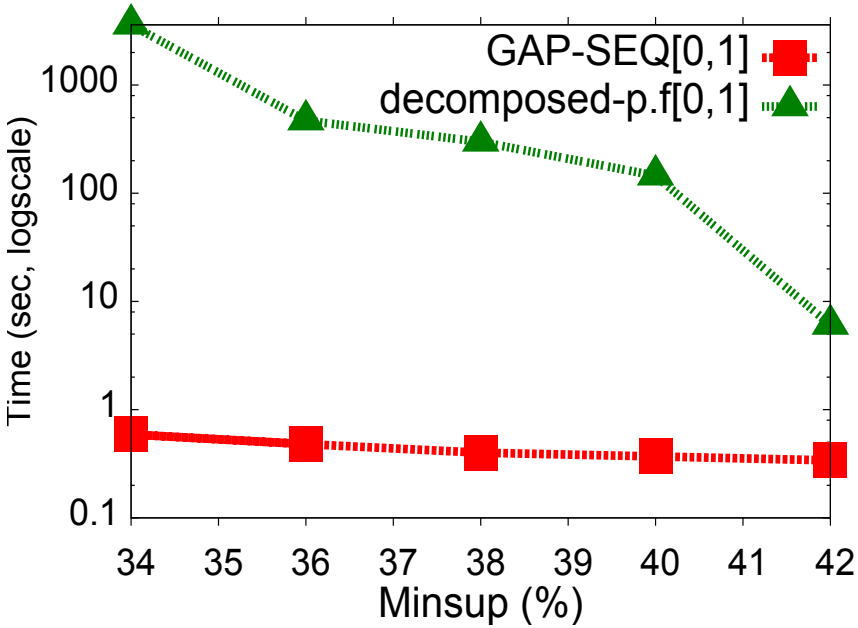


Résultats expérimentaux (3/5)

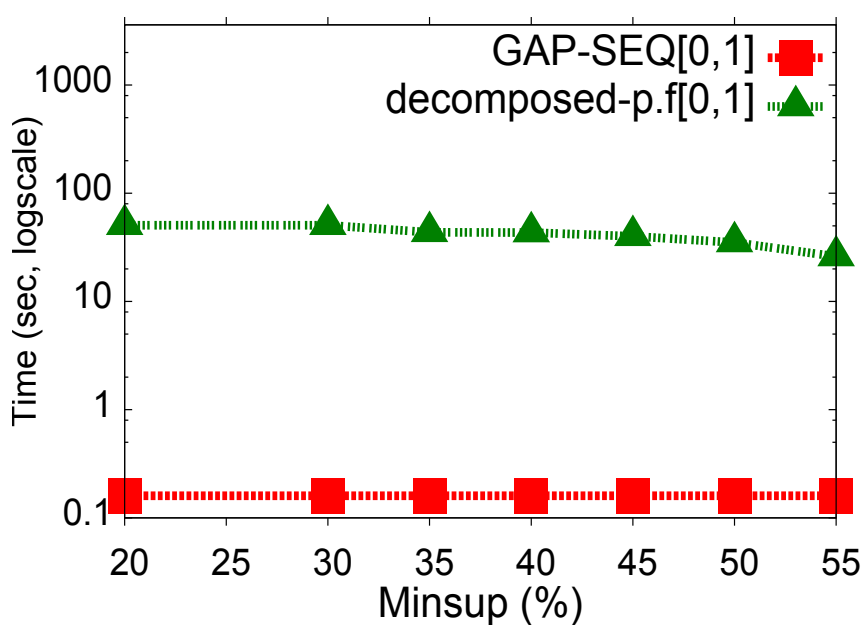
Comparaison avec une méthode CP : runtime

Varying the value of θ with the gap constraint $gap[0, 1]$

FIFA

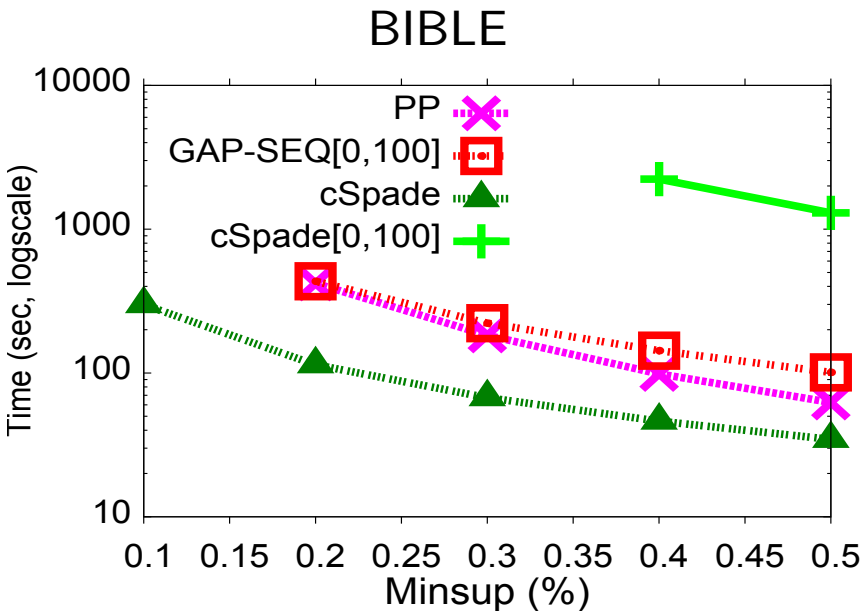


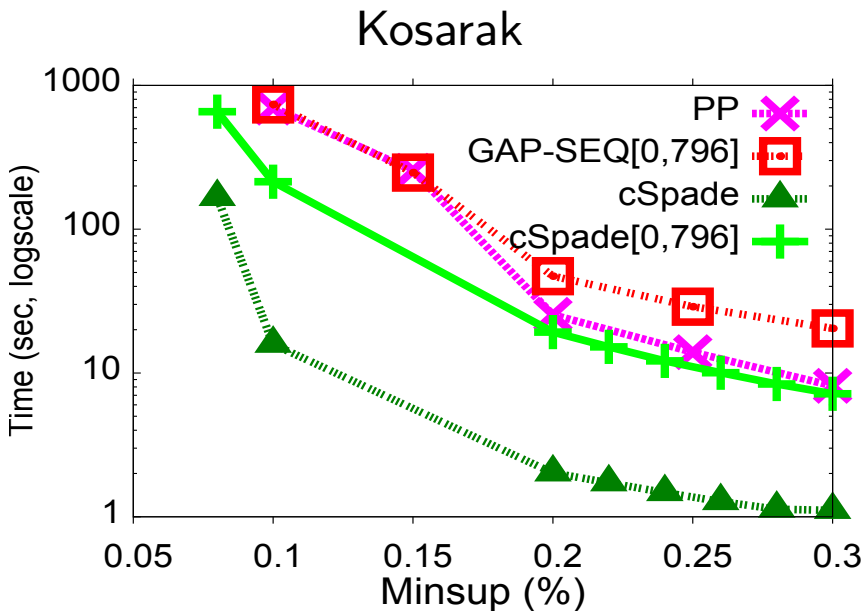
Leviathan

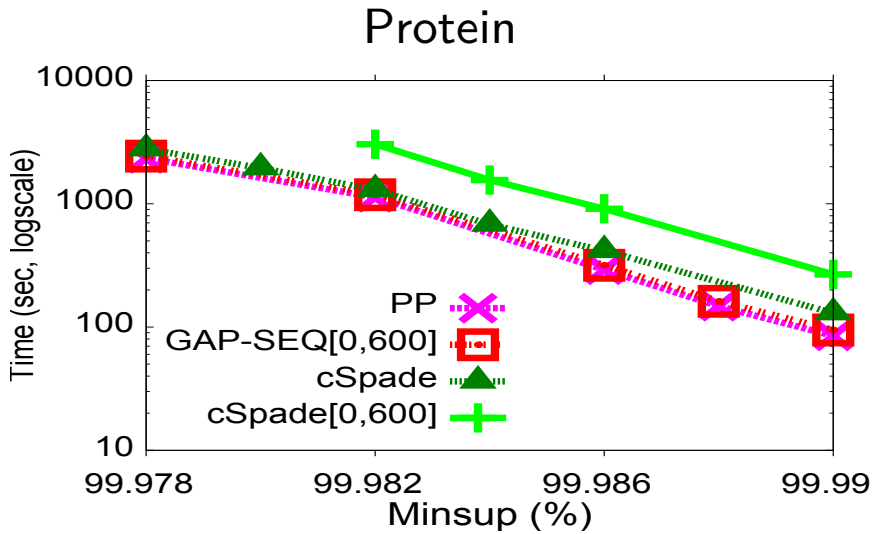


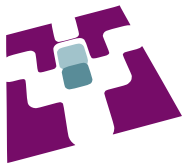
Résultats expérimentaux (4/5)

Comparaison avec PREFIX-PROJECTION : runtime









Résultats expérimentaux (5/5)

Ajout de contraintes supplémentaires

minsup	#PATTERNS		CPU times (s)		#PROPAGATIONS		#NODES	
	gap	gap+size+item	gap	gap+size+item	gap	gap+size+item	gap	gap+size+item
1 %	14032	1805	19.34	16.83	28862	47042	17580	16584
0.5 %	48990	6659	43.46	34.6	100736	163205	61149	58625
0.4 %	72228	10132	55.66	43.47	148597	240337	90477	87206
0.3 %	119965	17383	79.88	59.28	246934	398626	151280	146601
0.2 %	259760	39140	143.91	100.09	534816	861599	329185	321304
0.1 %	963053	153411	539.57	379.04	1986464	3186519	1236340	1219193

Bilan

Point de vue PPC

- plusieurs contraintes globales pour la fouille de séquences
- filtrage exploitant des propriétés bien connues en fouille

Contraintes	PP	SPADE	cSPADE	SPAM	SPIRIT	BIDE	GAP-BIDE	PrefixSpan
Fréquence minimale	✓	✓	✓	✓	✓	✓	✓	✓
Item	✓							
Taille	✓		✓					
Gap	✓		✓				✓	
Expressions régulières	✓				✓			
Fermeture	✗					✓	✓	
top-k	✓							

Point de vue modélisation et résolution

- cadre unifié pour "composer" plusieurs contraintes
- premières approches PPC qui passent véritablement à l'échelle

Perspectives

Perspectives

- Extension vers les séquences d'itemsets
- Implantation de la contrainte de fermeture
- Approches/contraintes pour l'extraction d'ensembles de motifs

Merci !

Special thanks to :

Nicolas Béchet (IRISA, France)

Amina Kemmar (University of Oran 1, Algeria)

Yahia Lebbah (University of Oran 1, Algeria)