

Sélection d'attributs de puce à ADN par essaim de particules

El-Ghazali Talbi
Laetitia Jourdan
Jose Nieto
Enrique Alba
Thomas Legrand

Paradiseo



Laboratoire d'Informatique
Fondamentale de Lille

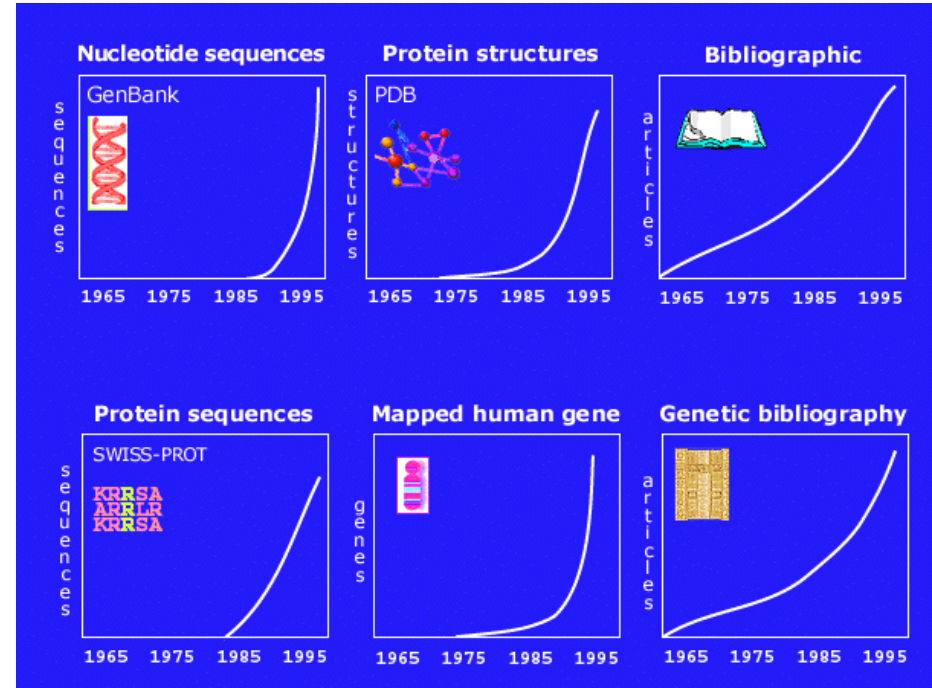


Plan

- Motivations biologiques
- La sélection d'attributs
- PSO binaire
- Résultats
- Conclusion et perspectives

Motivations

- Biologie : accroissement de la taille des données disponibles
 - Technologies haut débit (séquençage, génotypage, protéomique, transcriptome)



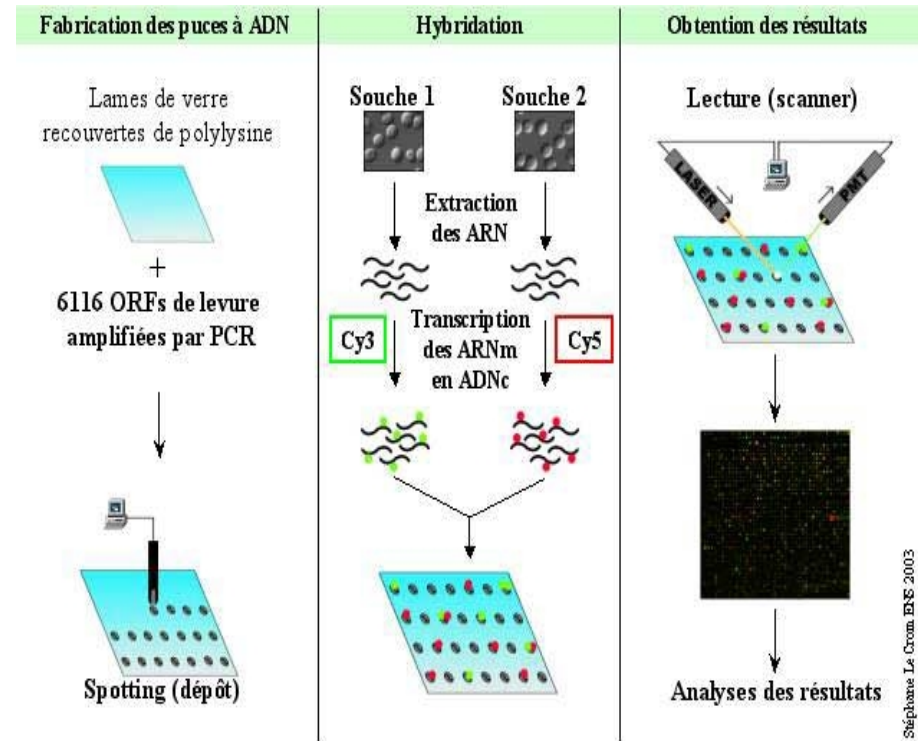
➡ **Besoin en Extraction de connaissances**

Puces à ADN (1)

- **Avantage principal des techniques “Puces à ADN”**
 - Permet l’analyse simultanée d’expressions de milliers de gènes dans une seule expérience
- **Processus “Puces à ADN”**
 - Arrayer
 - Expérience : Hybridation
 - Capture des images résultats
 - Analyse

Puces à ADN (2)

- Des robots alignent les ESTs (Expressed Sequence Tags) sur les lames de microscopes
- cellules mARN marquées par des tags fluorescents
- Liaison mARN - cADN exprimée (fluorescence) indique que le gène est actif

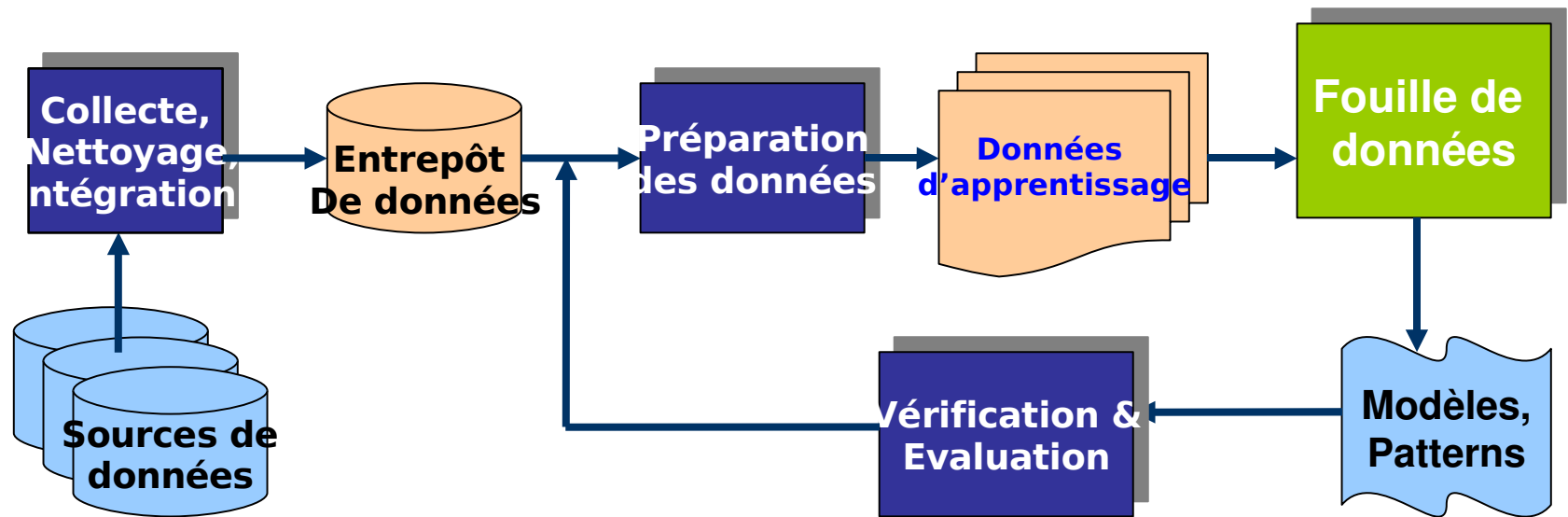


Données

- <http://sdmc.lit.org.sg/GEDatasets/Datasets.html>
- Grand nombre d'attributs/petit nombre d'observations

	# Attributs (gènes)	# Instances (Training/Test)
ALL-AML Leukemia	7129	72 (38/34)
Breast Cancer	24481	97 (78/19)
Colon Tumor	2000	62
Ovarian Cancer	15154	253
Lung Cancer	12533	181 (32/149)

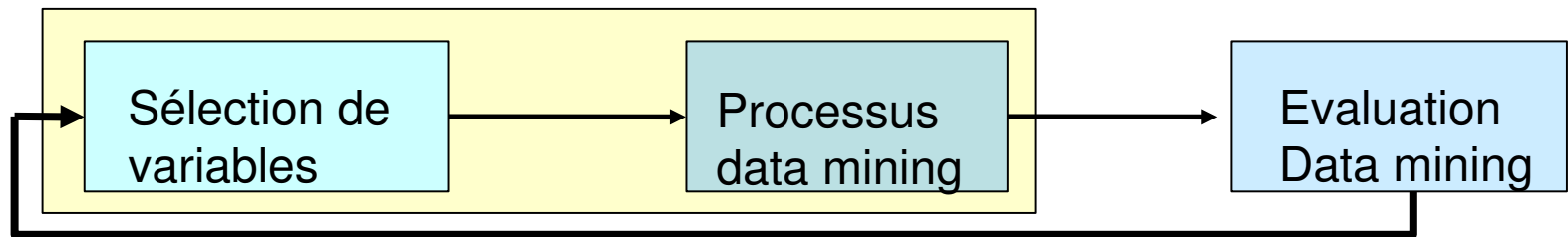
Extraction de connaissances



Sélection d'attributs (SA)

- **Objectif** : Réduire le nombre d'attributs considérés pour augmenter la précision de la prédiction et/ou réduire le temps de traitement des données, **Complexité** $\sum_{i=0}^M C_N^i$
- **2 Types de SA** :
 - Sélection par filtrage (filter) : la sélection se fait avant et indépendamment du processus de DataMining
 - Sélection par emballage (wrapper) : la sélection de variables englobe le processus de DM (ici classification)
 - Critère de sélection : performances du modèle appris
 - Coût de calcul très élevé car boucle sur le processus d'apprentissage
- **Classification**
 - Elle permet de **prédire** si un élément est membre d'un groupe ou d'une catégorie donné.
 - **Apprentissage supervisé** : classes connues à l'avance
 - Pb : qualité de la classification (taux d'erreur, accuracy)

Schéma Global de l'approche



Pour chaque évaluation

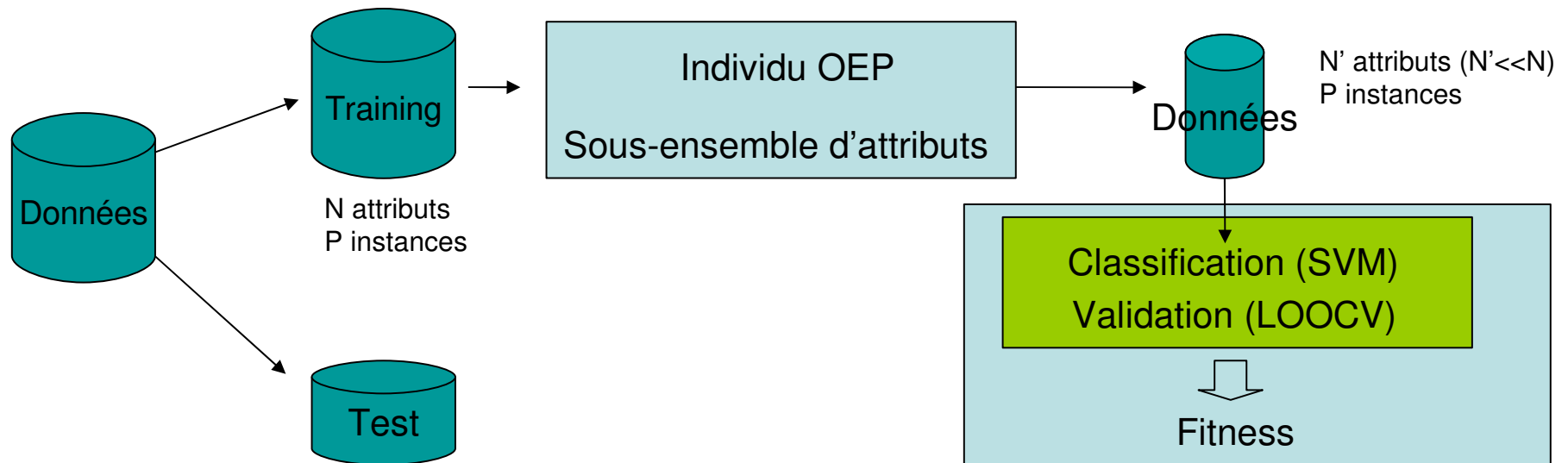


Schéma du Binary PSO

Algorithm 1 Binary PSO for feature selection.

```
1:  $S \leftarrow \text{SwarmInitialization}()$ 
2: while not stop condition do
3:   for  $i = 1$  to  $\text{size}(S)$  do
4:     evaluate each particle  $x_i$  of the swarm  $S$ 
5:     if  $\text{fitness}(x_i)$  is better than  $\text{fitness}(pBest_i)$  then
6:        $pBest_i \leftarrow x_i$ ;  $\text{fitness}(pBest_i) \leftarrow \text{fitness}(x_i)$ 
7:     end if
8:     if  $\text{fitness}(pBest_i)$  is better than  $\text{fitness}(g_i)$  then
9:        $g_i \leftarrow pBest_i$ ;  $\text{fitness}(g_i) \leftarrow \text{fitness}(pBest_i)$ 
10:    end if
11:  end for
12:  for  $i = 1$  to  $\text{size}(S)$  do
13:     $v_i \leftarrow \omega \cdot v_i + \varphi_1 \cdot \text{rand}_1 \cdot (pBest_i - x_i) + \varphi_2 \cdot \text{rand}_2 \cdot (g_i - x_i)$ 
14:     $x_i \leftarrow \text{update\_position}(x_i, v_i)$  // Equation 4
15:  end for
16: end while
17: Output: best solution found
```

Application OEP/SA

- **Codage** : Binaire : 1 indique que le gène est sélectionné, 0 indique que le gène n'est pas sélectionné
- **Evaluation** : Classification par SVM [Vapnik] (accuracy) + Leave-one-out cross-validation (LOOCV) + prise en compte du nombre d'attributs sélectionnés avec pondération (agrégation)

$$fitness(x_i) = \alpha \times \left(\frac{100}{accuracy} \right) + \beta \times (features)$$

Application OEP/SA (2)

- **Mise à jour du mouvement :**

$$x_{iD}(new) = \begin{cases} x_{iD}(new), & (0 < v_{iD} \leq a) \\ pBest_{iD}, & (a < v_{iD} \leq \frac{1+a}{2}) \\ g_{iD}, & (\frac{1+a}{2} < v_{iD} \leq 1) \end{cases}$$

→ a généré aléatoirement dans [0.1] pour chaque composante.

Résultats

- Comparaison de notre PSO avec la littérature et avec un AG implémenté sous ParadisEO (même fonction d'évaluation)
- Initialisation pour assurer des individus avec différents nombres d'attributs sélectionnés

PSO		GA	
Parameter	Value	Parameter	Value
Swarm size	40	Population size	40
Number of generations	50	Number of generations	50
Neighborhood size	8	Probability of crossover	0.9
Inertia (min,max)	(0.4,0.9)	Probability of mutation	0.1
Velocity (min,max)	(0.0,1.0)	-	-

Résultats

Taux de classification

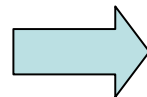
Nombre de gènes

Dataset	PSO_{SVM}	GA_{SVM}	[12]	[6]	[27]	[7]	[28]	[29]	[23]
Luekemia	99.33(3)	99.79(4)	100(25)	-	100(4)	100(2)	87.55(4)	-	-
Breast	90.64(4)	89.02(4)	-	-	-	-	79.38(67)	-	-
Colon	100(3)	100(4)	99.41(10)	94.12(37)	97.0(7)	98.0(4)	93.55(4)	85.48(-)	94.00(4)
Lung	100(4)	97.38(3)	-	-	-	-	98.34(6)	-	-
Ovarian	100(4)	100(3)	-	-	-	-	-	99.21(75)	-
Prostate	100(4)	100(4)	-	88.88(20)	-	-	-	-	-

Moyenne sur 30 exécutions vs best known

Résultats

Dataset	PSO_{SVM}			GA_{SVM}		
	Best	Mean	Std Dev.	Best	Mean	Std Dev.
Leukemia	100(3)	99.33(3)	0.9439	100(4)	99.79(4)	0.6797
Breast	100(4)	90.64(4)	7.3075	100(4)	89.02(4)	7.4696
Colon	100(3)	100(3)	0.0000	100(4)	100(4)	0.0000
Lung	100(4)	100(4)	0.0000	100(3)	97.38(3)	4.1684
Ovarian	100(4)	100(4)	0.0000	100(2)	100(3)	0.0000
Prostate	100(3)	100(4)	0.0000	100(3)	100(4)	0.0000



Robustesse

Gènes sélectionnés

- Les gènes sont donnés aux biologistes.

Dataset	PSO_{SVM}		GA_{SVM}	
Luekemia	100(3)	<i>K01383_at, U03056_at, J04130_s_at</i>	100(4)	<i>L40379_at, S85963_at U83192_at, Z49099_at</i>
Breast	100(4)	<i>Contig49744_RC, Contig26884_RC Contig25936_RC, Contig13846_RC</i>	100(4)	<i>NM_002593, NM_012084 NM_003603, NM_003743</i>
Colon	100(3)	<i>H64398, H73758 U27699</i>	100(4)	<i>T95318, U31383 H89092, U19261</i>
Lung	100(4)	<i>31681_at, 32461_f_at 728_at, 829_s_at</i>	100(3)	<i>33762_r_at, 34648_at 36226_r_at</i>
Ovarian	100(4)	<i>MZ141.59378, MZ4801.6282 MZ6158.6231, MZ14383.58</i>	100(2)	<i>MZ1154.6306 MZ2653.8464</i>
Prostate	100(3)	<i>35404_at, 40076_at 393_s_at</i>	100(3)	<i>35935_at, 39801_at 40069_at</i>

Conclusion

- Travail utilisant les OEP en binaire
- OEP : très bon résultats pour la sélection d'attributs par rapport à la littérature
- Très peu de travaux en datamining avec les OEP

Perspectives

- Test des « geometrics PSO » [A. Moraglio, C. D. Chio, and R. Poli, “Geometric Particle Swarm Optimization,” in *10th European conference on Genetic Programming (EuroGP 2007)*]
- OEP multiobjectif :
 - Taux de classification
 - Nombre d'attributs
- Test d'autres classifieurs (K-NN, ...)
- Parallélisme de l'évaluation