

---

# Bases de données INductive et GénOmique

J-F. Boulicaut (LIRIS, Lyon), B. Crémilleux (GREYC, Caen),  
O. Gandrillon (CGMC, Lyon), F. Jacquenet (EURISE, St-Etienne)

<http://www.info.unicaen.fr/~bruno/bingo/>

Projet BINGO- ACI MD 46, 2004-2007

# BINGO : un projet de recherche en informatique et pluridisciplinaire

---

- **CGMC** - CNRS UMR 5534 (Lyon) : multiples problématiques biologiques sur l'analyse de données génomiques et transcriptomiques, préparation des données, rôle majeur dans la validation des résultats, concepts, méthodes et outils développés
- **EURISE** - EA 3721 (Saint-Etienne) : inférence grammaticale et extraction de connaissances dans des données séquentielles, apprentissage automatique et fouille de textes
- **GREYC** - CNRS UMR 6072 (Caen) : extraction sous contraintes et usages multiples des motifs ensemblistes, classification et clustering, extraction d'information dans des textes et linguistique
- **LIRIS** - CNRS UMR 5205 (Lyon) : cadre des bases de données inductives, extraction sous contraintes de motifs ensemblistes et séquentiels, fouille de données du transcriptome

# Buts de BINGO

---

Recherche dans le cadre des bases de données inductives

- optimisation de l'*évaluation de requêtes inductives complexes*
- *usages multiples des motifs* pour la construction de modèles (e.g., classifieurs, soft-clustering)
- *utilisation de ressources textuelles* : optimiser le post-traitement et l'extraction de motifs à l'aide de ressources textuelles
- *prototypage d'un langage de requêtes* pour la mise en œuvre de scénarios ECBD en biologie moléculaire

➔ problèmes ouverts étudiés via des applications à la génomique

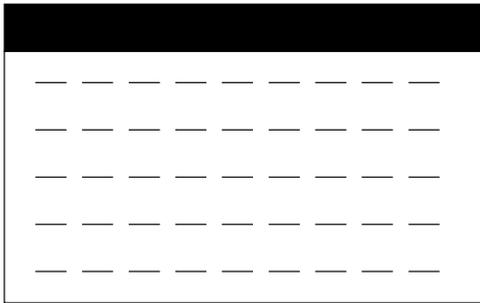
# Exemples de requêtes inductives

---

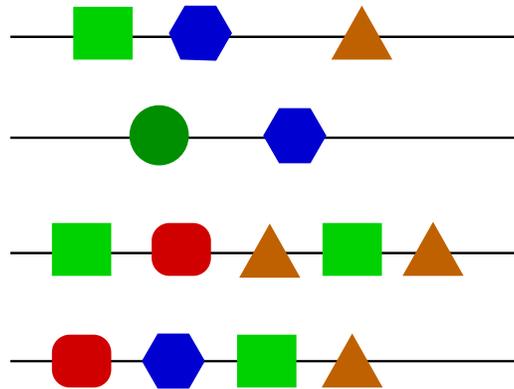
- **R1** : rechercher tous les ensembles de gènes *fréquemment sur-exprimés dans le sous-ensemble  $D_1$  et non fréquemment sur-exprimés dans le sous-ensemble  $D_2$*   
↳ motifs dits “émergents”
- **R2** : rechercher tous les ensembles de gènes *fréquemment sur-exprimés et qui partagent au moins un motif dans leurs séquences promotrices et dont les annotations dans les ressources textuelles associées dépassent un niveau de similarité donné*

# Des sources de données multiples

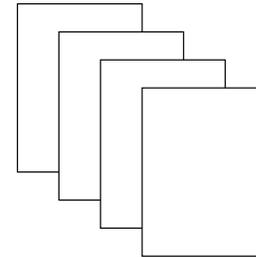
expression de gènes



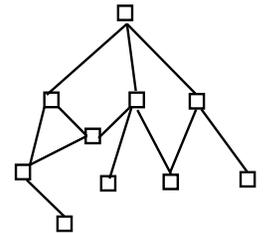
séquences de promoteurs



ressources  
textuelles



ontologies



Données du transcriptome :

Serial Analysis of Gene Expression (SAGE)

*avantages des données SAGE :*

- exhaustivité
- comparaison directe de bibliothèques
- capacité à isoler de nouveaux gènes

# Des motifs locaux aux motifs globaux et modèles

*Exemples :*

- de bi-ensembles à des bi-partitions
- de motifs émergents à des classifieurs
- de règles de caractérisation à l'interprétation de clusters

	$g_1$	$g_2$	$g_3$	$g_4$	$g_5$
$o_1$	X		X	X	
$o_2$		X			X
$o_3$	X		X	X	
$o_4$			X	X	
$o_5$	X	X			X
$o_6$		X			X
$o_7$					X

*Sur les méthodes :*

- *motifs locaux* : méthode complète
- *modèle* : utilisation d'heuristique

# Un cadre pour le bi-clustering

CGMC/LIRIS

	$g_1$	$g_2$	$g_3$	$g_4$	$g_5$
$o_1$	X		X	X	
$o_2$		X			X
$o_3$	X		X	X	
$o_4$			X	X	
$o_5$	X	X			X
$o_6$		X			X
$o_7$					X

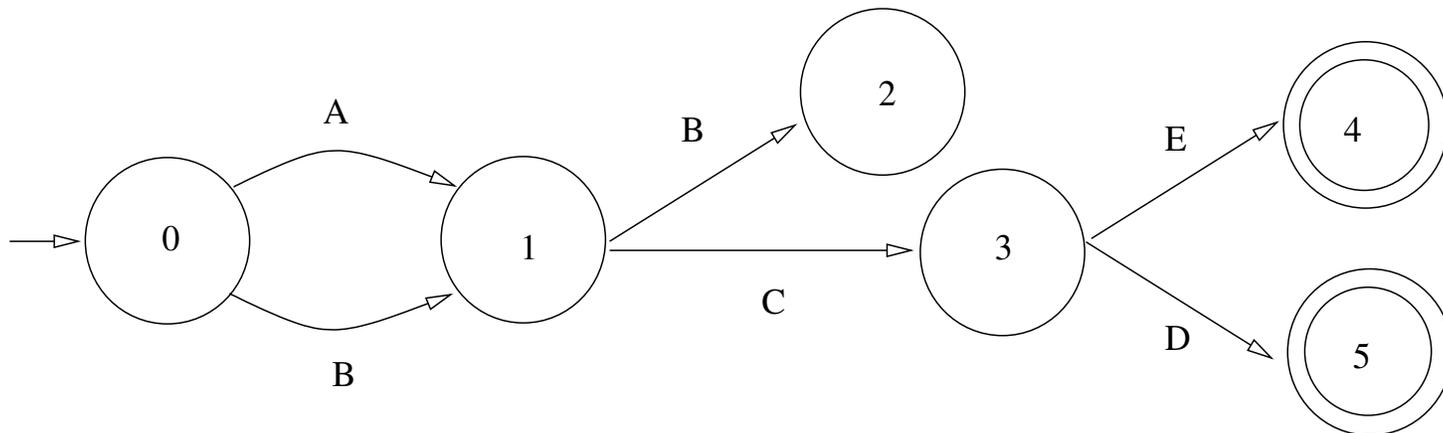
- groupement de motifs locaux (e.g., bi-ensembles tolérants aux erreurs)
- *motif global* : une bi-partition (ici ■ et ■), chevauchement entre bi-clusters possible
- solveur CDK-MEANS

# Fouiller les automates

## ... plutôt que les données

EURISE

données séquentielles  $\rightsquigarrow$  automates  $\rightsquigarrow$  fouille d'automates



$$E_+ = \{ACE, BCD\} \quad E_- = \{AB, BB\}$$

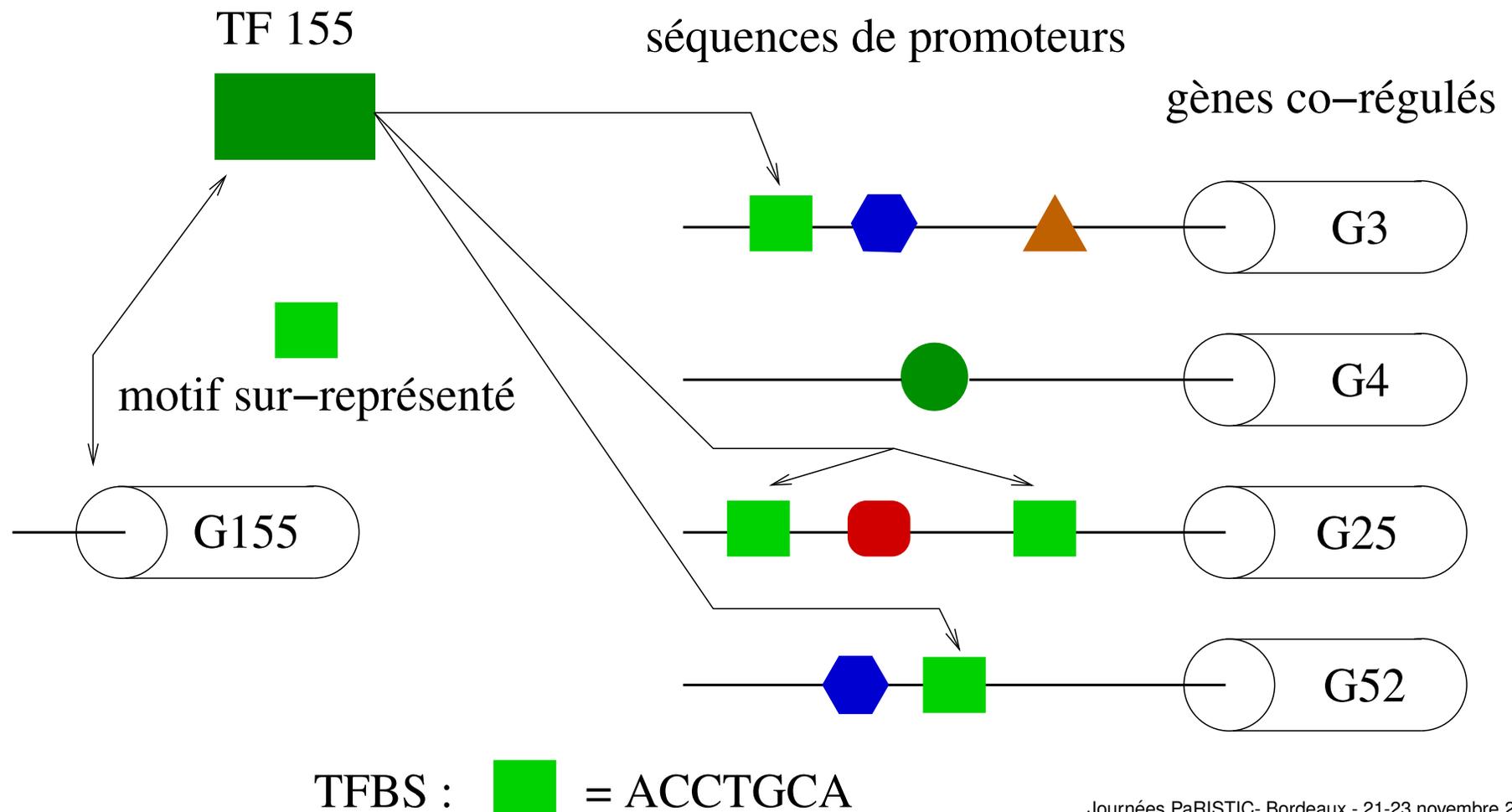
*Intérêts :*

- “nouveaux motifs” (e.g., ACD et BCE) non ou insuffisamment couverts par les données
- contraintes (préfixe, statistiques)

# Perspectives : construction de réseaux d'expression génique

CGMC/EURISE/LIRIS

De l'analyse des données d'expression de gènes à la construction de réseaux d'expression génique



# Perspectives : extraction de motifs sous contraintes textuelles

CGMC/EURISE/GREYC

