Bases de données INductives 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

Plan

- introduction au projet BINGO
- *focus sur quelques résultats*
  - extraction de motifs locaux
  - construction de motifs globaux
- *focus sur quelques perspectives*
  - des données d'expression de gènes à la construction de réseaux d'expression génique
  - optimiser le post-traitement et l'extraction de motifs à l'aide de ressources textuelles

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

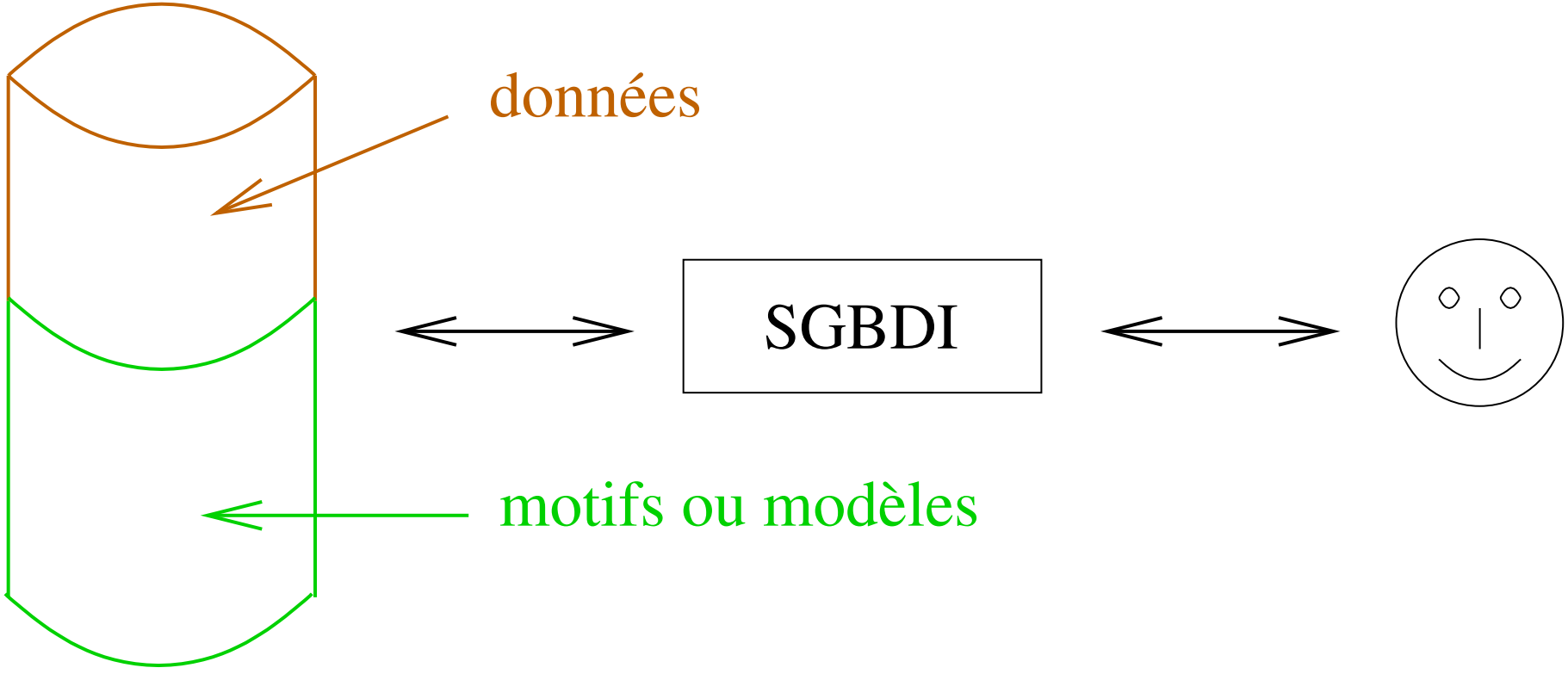
Recherche dans le cadre des bases de données inductives

- 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

Cadre des bases de données inductives

➠ processus ECBD comme séquences de requêtes sur données et motifs ou modèles



Exemples de requêtes inductives

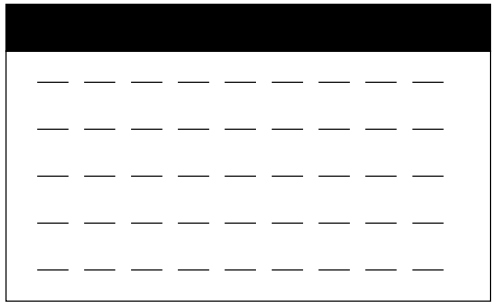
- **R1 :** rechercher tous les ensembles de gènes *fréquemment sur-exprimés dans le sous-ensemble* D_1 et *non fréquemment sur-exprimés dans le sous-ensemble* D_2

  ➥ motifs dits “émergents”

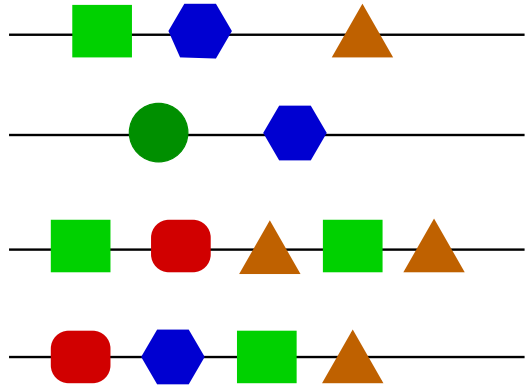
- **R2 :** rechercher tous les ensembles de gènes *fréquemment sur-exprimés* et qui *partagent au moins un motif dans leurs séquences promotrices* et dont *les annotations dans les ressources textuelles associées dépassent un niveau de similarité donné*

Des sources de données multiples

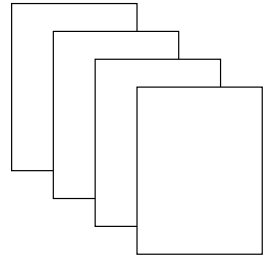
expression de gènes



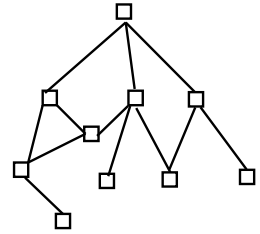
séquences de promoteurs



ressources textuelles



ontologies



Données du transcriptome :

➠ Serial Analysis of Gene Expression (SAGE)

avantages des données SAGE :

- exhaustivité
- comparaison directe de librairies
- 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

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

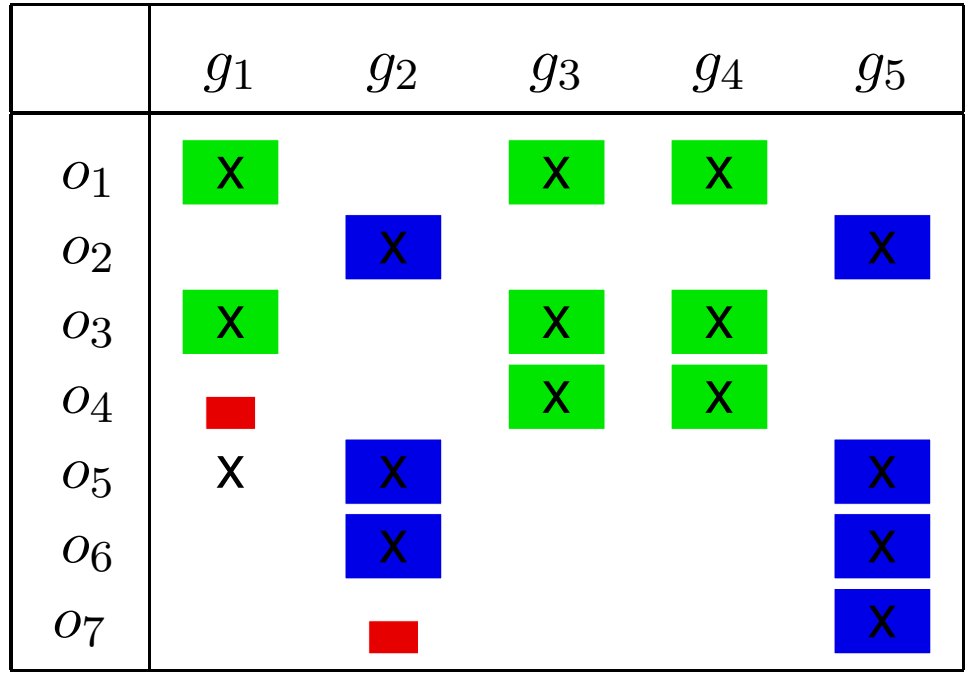
Sur les méthodes :

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

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



Sur les méthodes :

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

Recherche de bi-ensembles tolérants aux erreurs

CGMC/LIRIS

- concept formel

	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

Recherche de bi-ensembles tolérants aux erreurs

CGMC/LIRIS

- concept formel

	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

Recherche de bi-ensembles tolérants aux erreurs

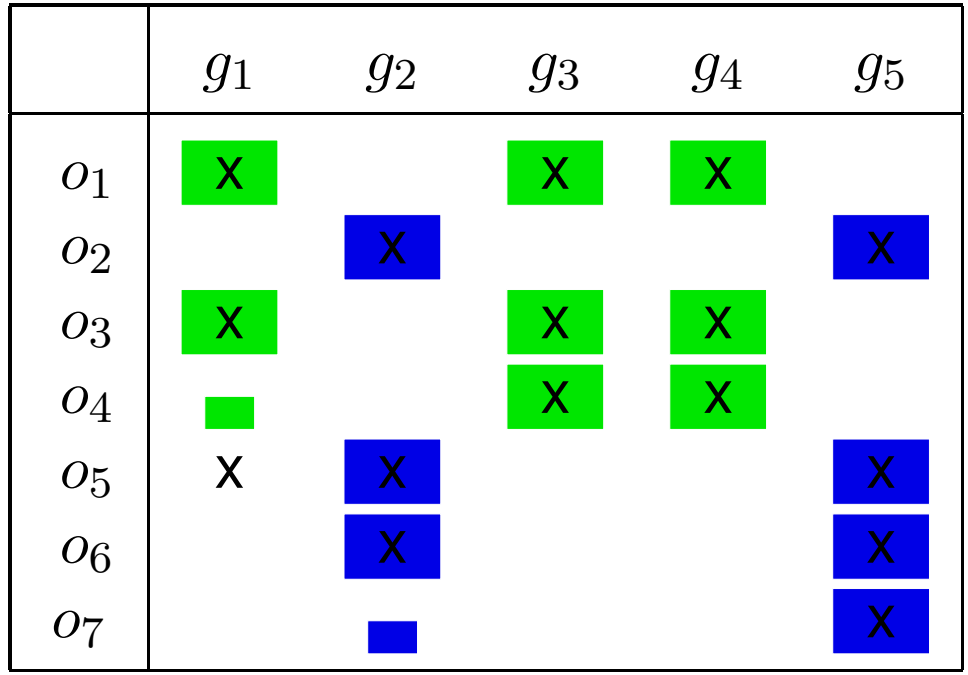
CGMC/LIRIS

	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

- bi-ensembles tolérants aux erreurs des données :
- solveur DR-MINER : bi-ensembles denses et pertinents

Un cadre pour le bi-clustering

CGMC/LIRIS



- 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

Règles de caractérisation dans les données génomiques

CGMC/GREYC

larges jeux de données : ➠ impossibilité d'extraire les règles de caractérisation δ-fortes

Solution :

- utilisation de l'extension associée à un motif d'attributs
- critère d'élagage propre aux règles de caractérisation δ-fortes (contrainte anti-monotone)

➥ solveur FTCMINER

Sur les données SAGE d'expression de gènes :

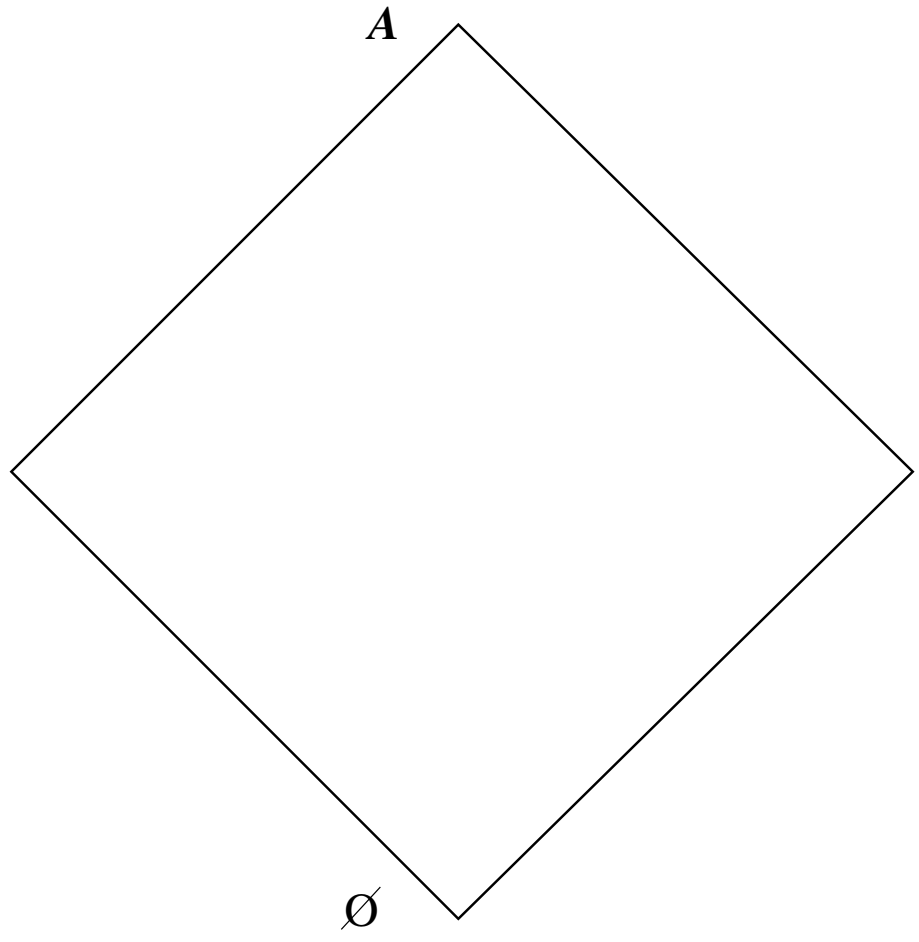
- des règles suggèrent un lien entre deux protéines et le cancer
- à raffiner suivant le type de cancer

Extraction sous contraintes : pousser des contraintes monotones et anti-monotones

- une *approche générique d'extraction de bi-ensembles* sous contraintes robustes aux erreurs (LIRIS)

- *"primitive-based framework"* :
  un cadre générique de contraintes (GREYC)

  Motifs virtuels pour automatiquement inférer des conditions d'élagage liées à la monotonie pour toute contrainte du cadre.

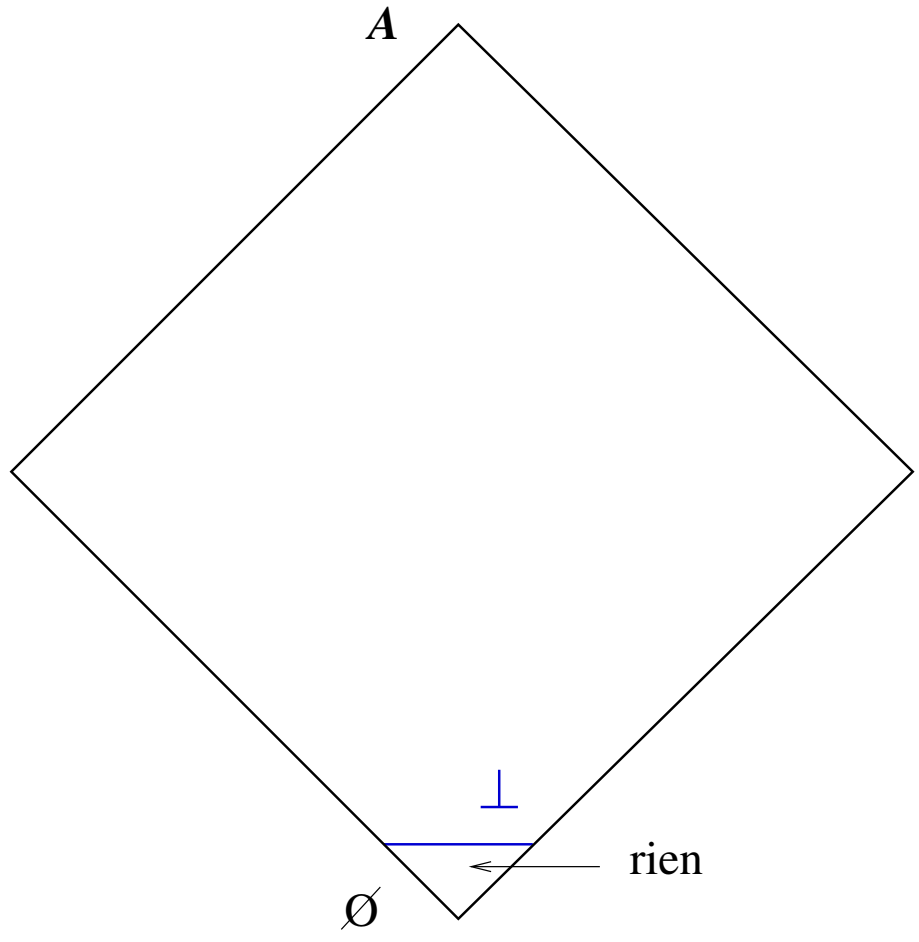


Extraction sous contraintes : pousser des contraintes monotones et anti-monotones

- une *approche générique d'extraction de bi-ensembles* sous contraintes robustes aux erreurs (LIRIS)

- *"primitive-based framework"* :
  un cadre générique de contraintes (GREYC)

  Motifs virtuels pour automatiquement inférer des conditions d'élagage liées à la monotonie pour toute contrainte du cadre.

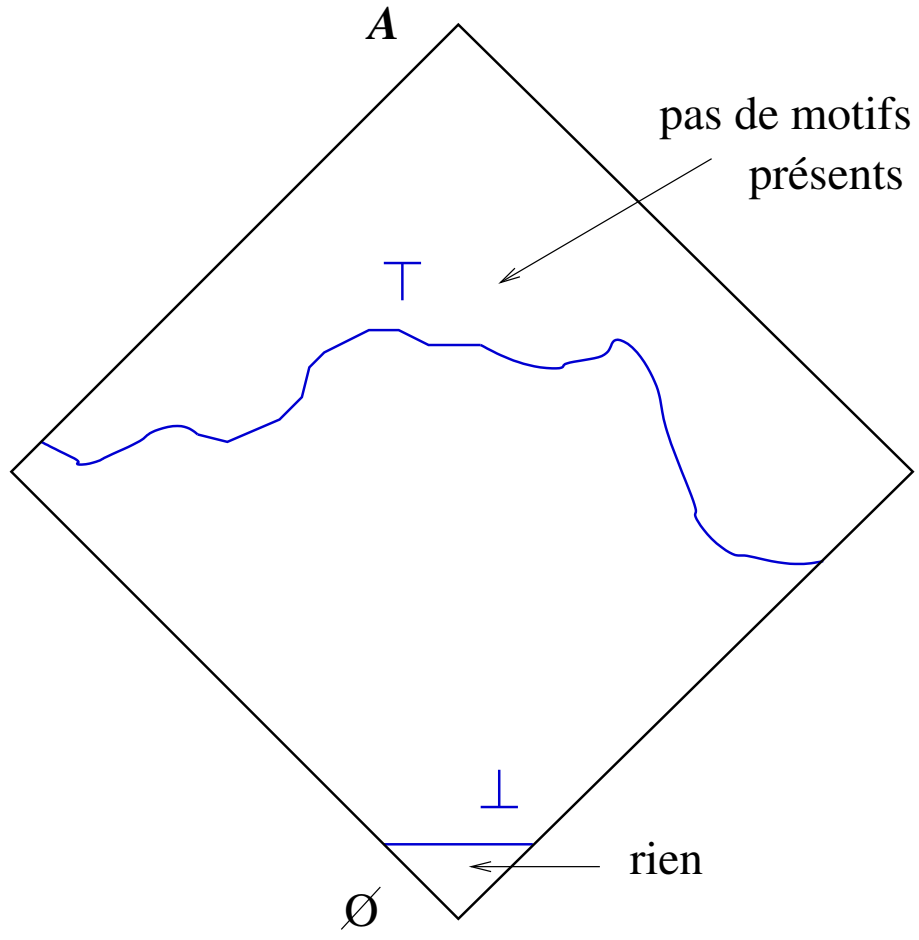


Extraction sous contraintes : pousser des contraintes monotones et anti-monotones

- une *approche générique d'extraction de bi-ensembles* sous contraintes robustes aux erreurs (LIRIS)

- *"primitive-based framework"* :
  un cadre générique de contraintes (GREYC)

  Motifs virtuels pour automatiquement inférer des conditions d'élagage liées à la monotonie pour toute contrainte du cadre.

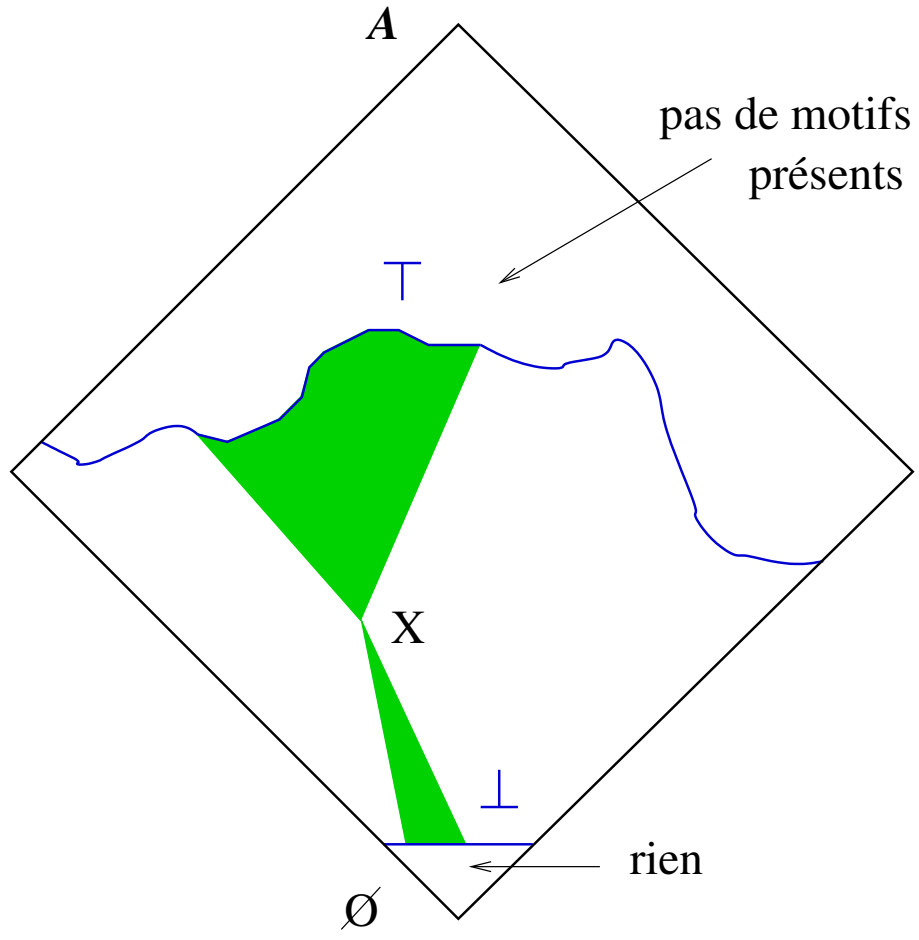


Extraction sous contraintes : pousser des contraintes monotones et anti-monotones

- une *approche générique d'extraction de bi-ensembles* sous contraintes robustes aux erreurs (LIRIS)

- *"primitive-based framework"* :
  un cadre générique de contraintes (GREYC)

  Motifs virtuels pour automatiquement inférer des conditions d'élagage liées à la monotonie pour toute contrainte du cadre.

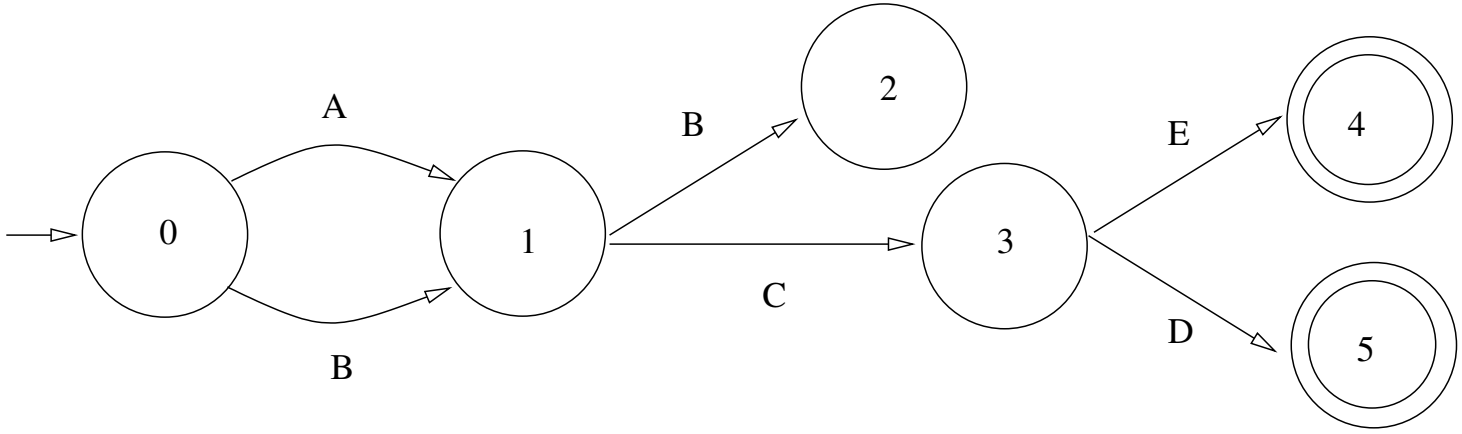


Fouiller les automates

…plutôt que les données

EURISE

données séquentielles ➠ automates ➠ fouille d'automates



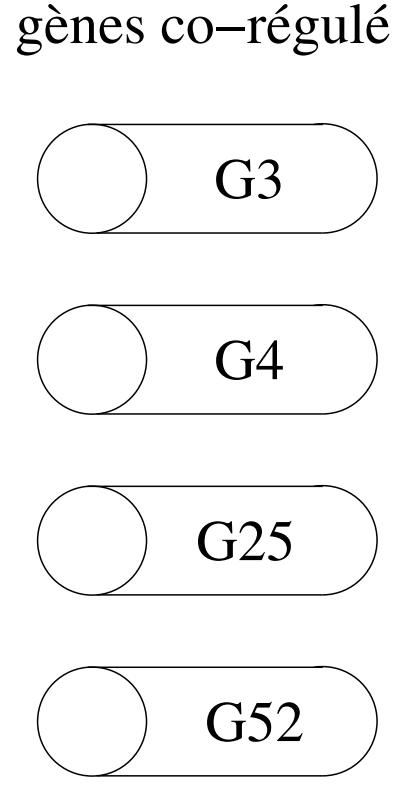
$E_+ = \{\text{ACE, BCD}\}$ $E_- = \{\text{AB, BB}\}$

Intérêts :

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

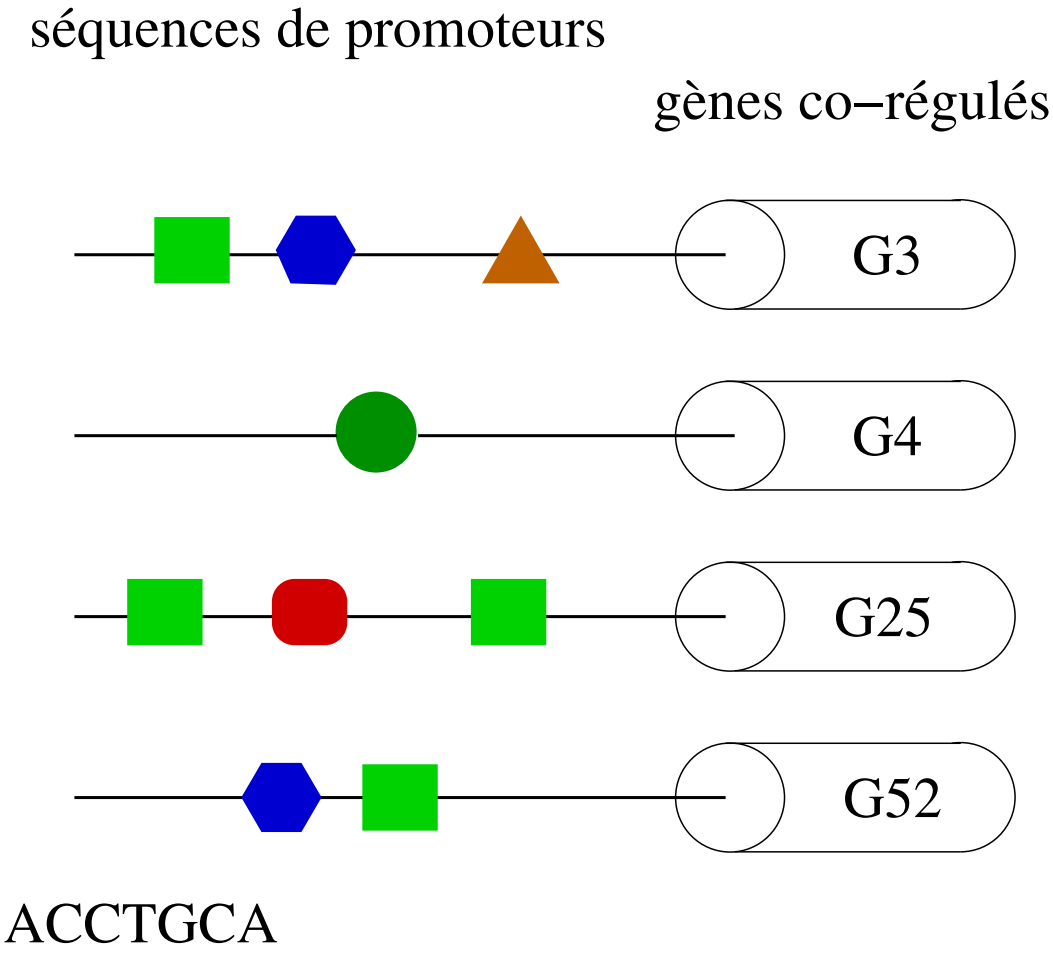
Perspectives : construction de réseaux d'expression génique

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



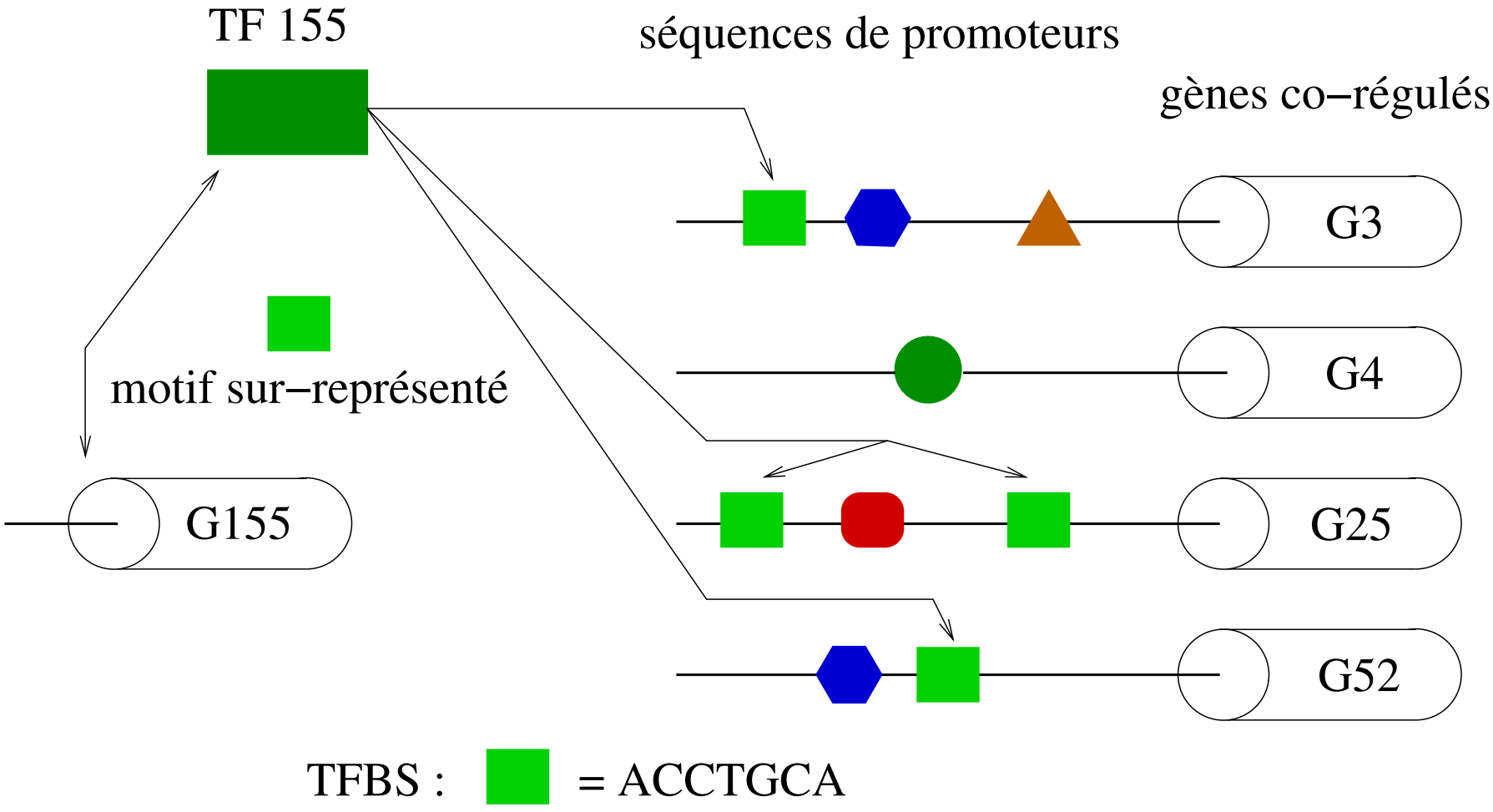
Perspectives : construction de réseaux d'expression génique

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



Perspectives : construction de réseaux d'expression génique

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

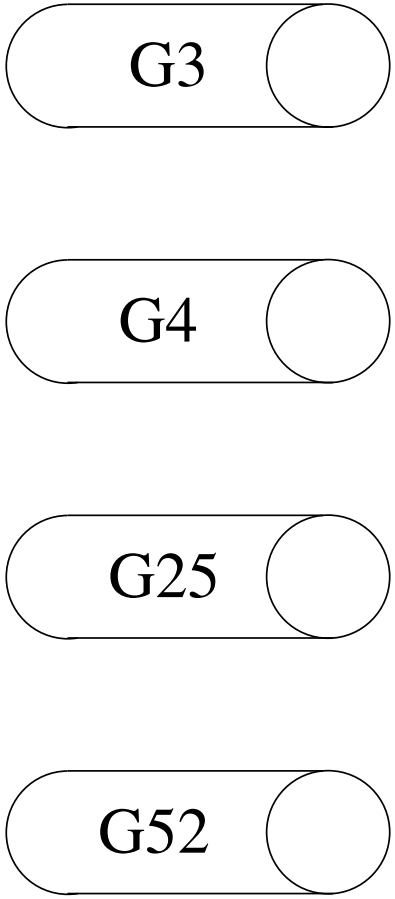


Perspectives : construction de réseaux d'expression génique

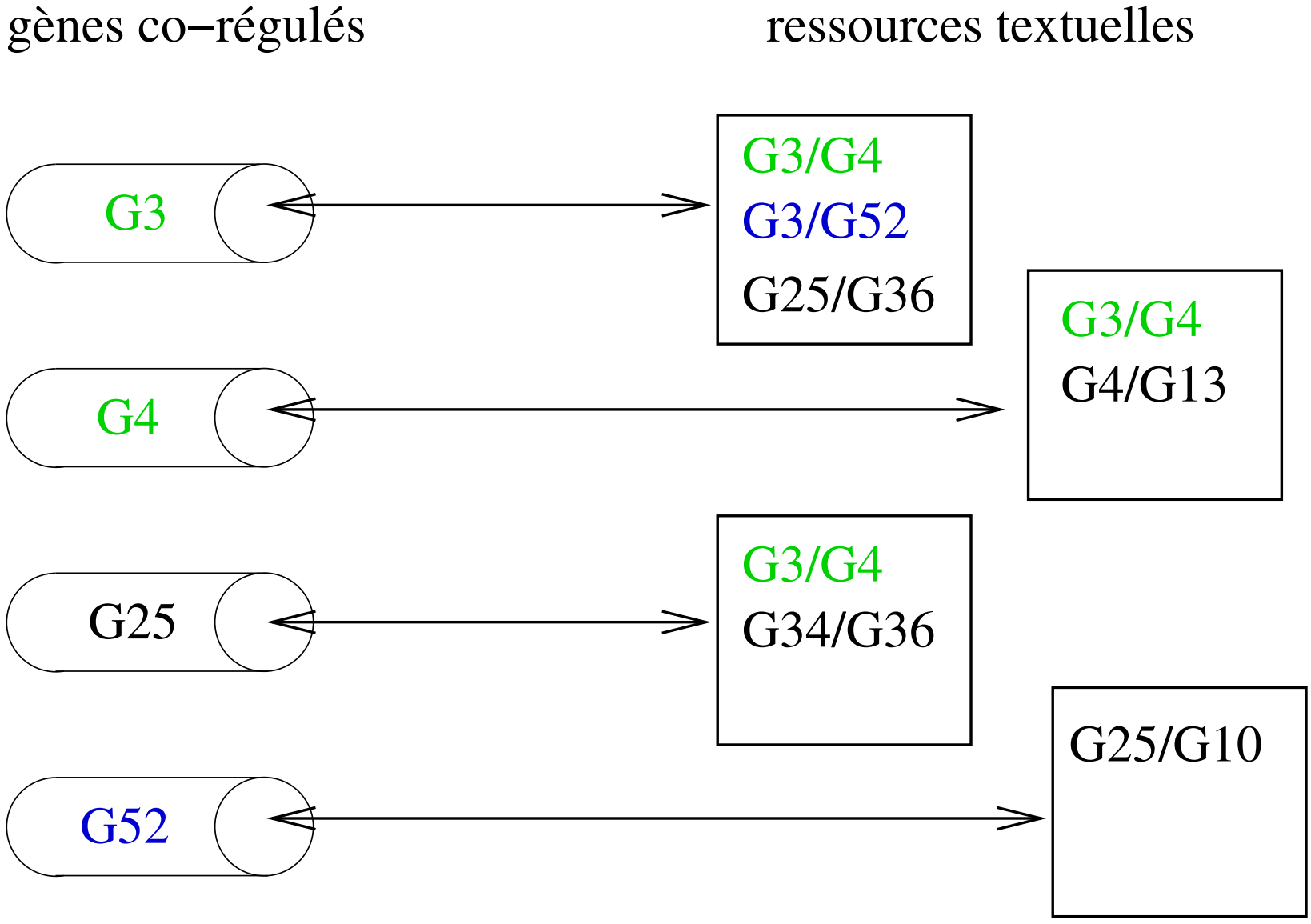
- CGMC : préparation des données d'expression de gènes, connaissances du domaine, validation des résultats
- EURISE : construction d'automates, fouille d'automates avec des techniques d'inférence statistique robustes au bruit
- LIRIS : extraction de motifs séquentiels sous contraintes dans de grands jeux de données (fouille des séquences de promoteurs)

Apport des ressources textuelles

gènes co–régulés



Apport des ressources textuelles



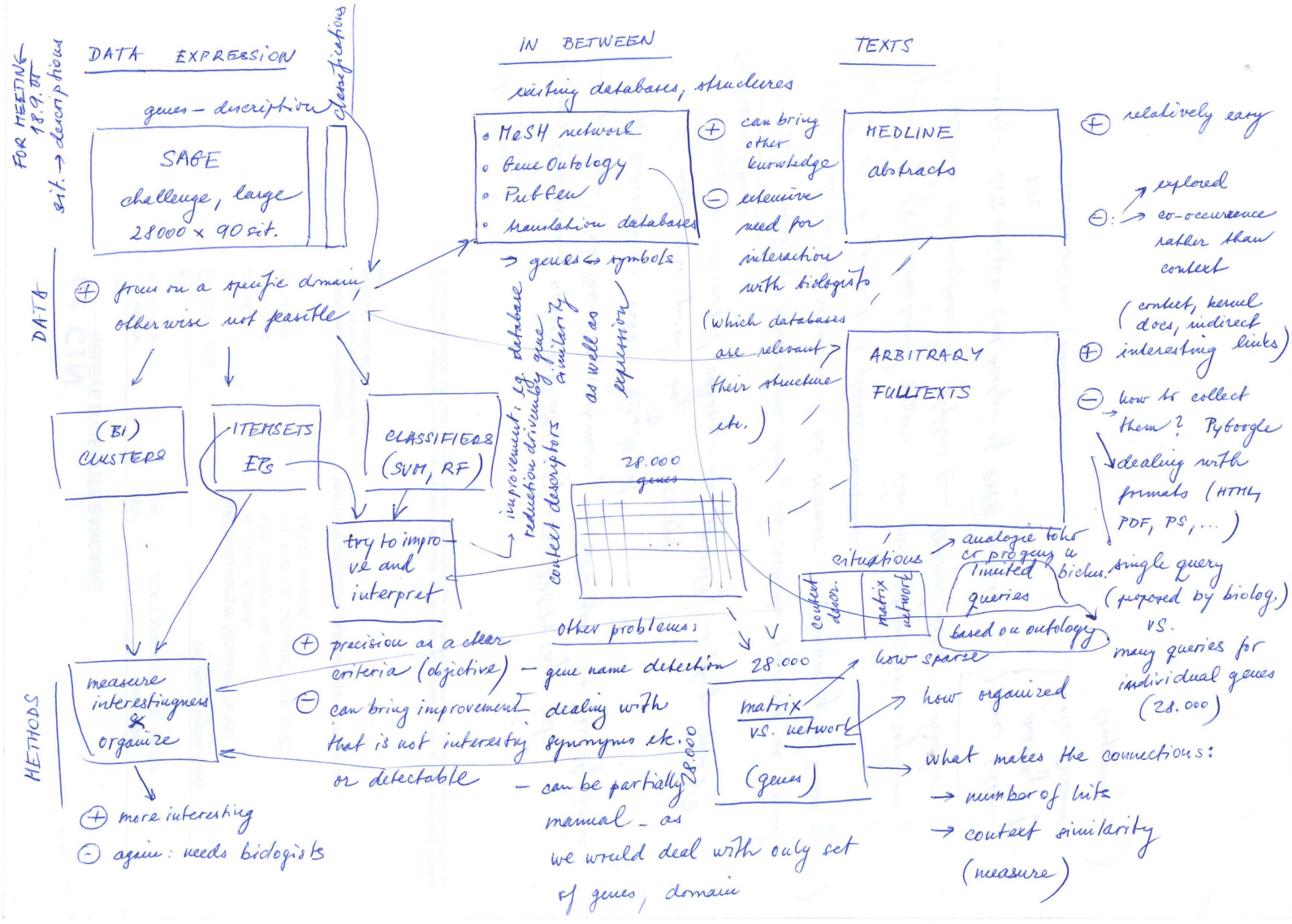
L'utilisation des ressources textuelles dans BINGO

But : optimiser le post-traitement et l'extraction de motifs à l'aide de ressources textuelles.

- *approche globale (macro-textuelle) :*
  - qualifier la façon dont on parle d'un objet
  - possibilité de prise en compte de la structure du document (hiérarchie de descripteurs) :
    ➥ apport de descripteurs de style pour la caractérisation du type d'articles
- *approche locale (intra-phrastique) :* repérer des entités nommées à partir d'une analyse locale

démarche : combiner ces approches sur le texte avec des méthodes de fouille de données

Perspectives : utilisation des ressources textuelles dans BINGO



Perspectives : extraction de motifs sous contraintes textuelles

CGMC/EURISE/GREYC