Apprentissage de la structure de réseaux bayésiens. Application aux données de génétique-génomique.

Jimmy Vandel

Directeurs de thèse : *Brigitte Mangin & Simon de Givry*

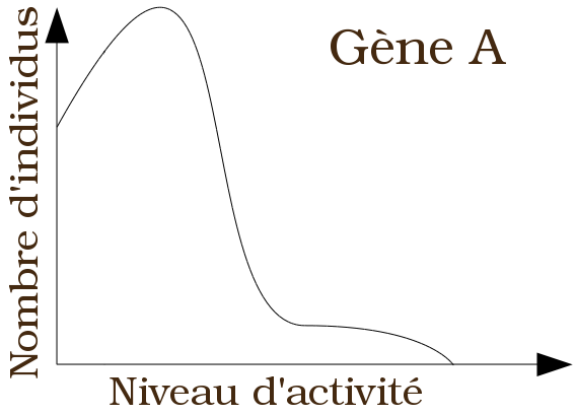
Université Paul Sabatier
TOULOUSE III

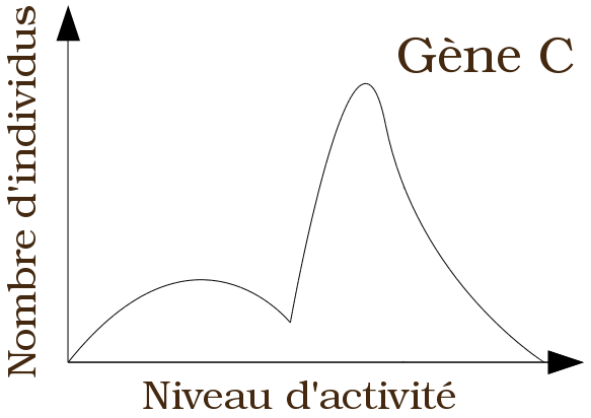
École Doctorale Mathématiques, Informatique et Télécommunications de Toulouse

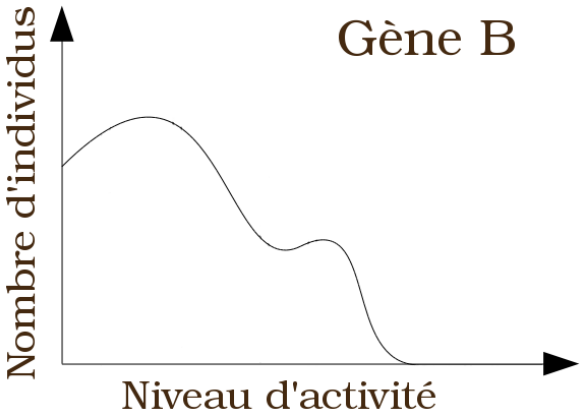
Unité BIA Toulouse
Biométrie et Intelligence Artificielle

Motivation biologique

- Mesure pour un ensemble d'individus, l'activité de différents gènes (par exemple le niveau d'expression)

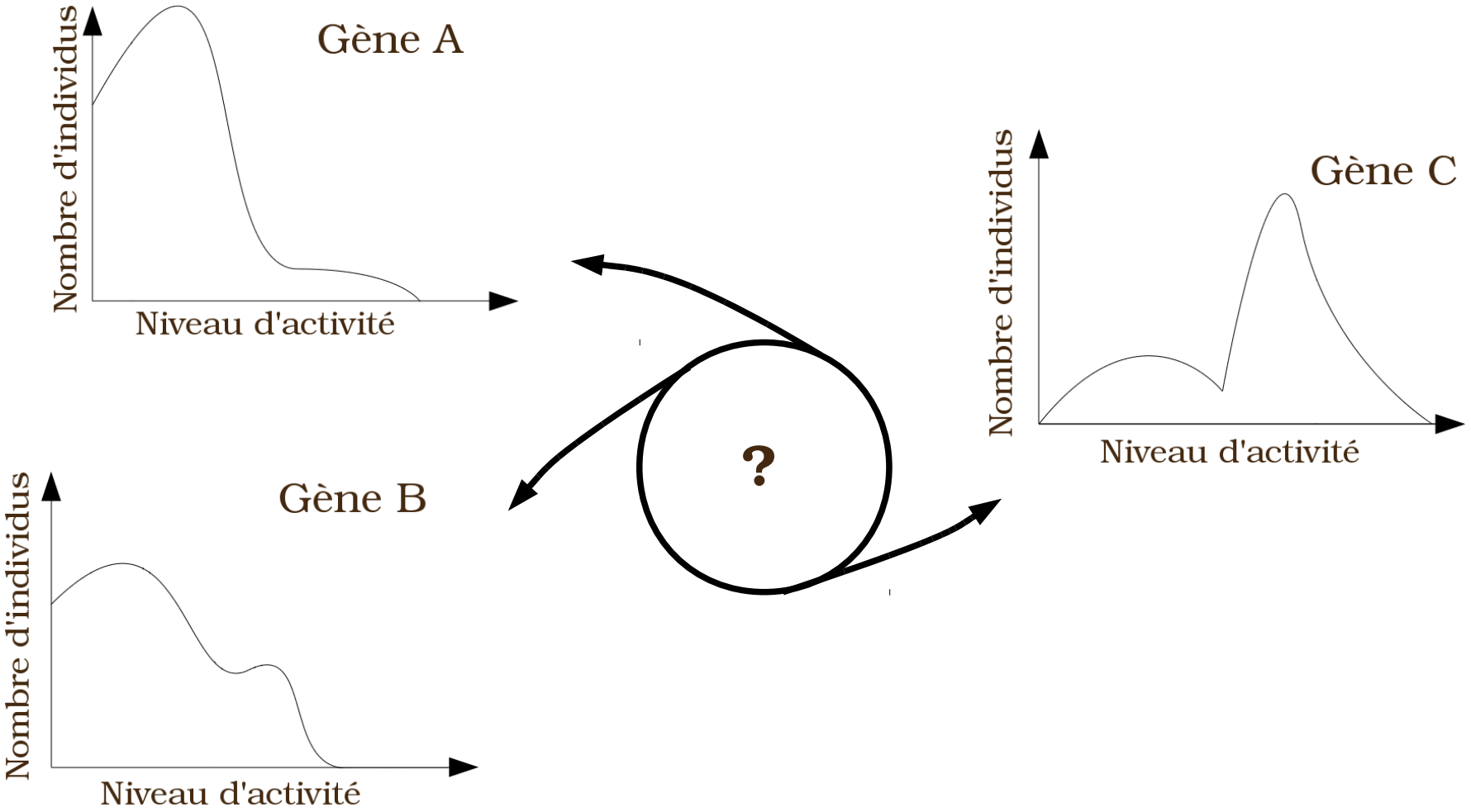






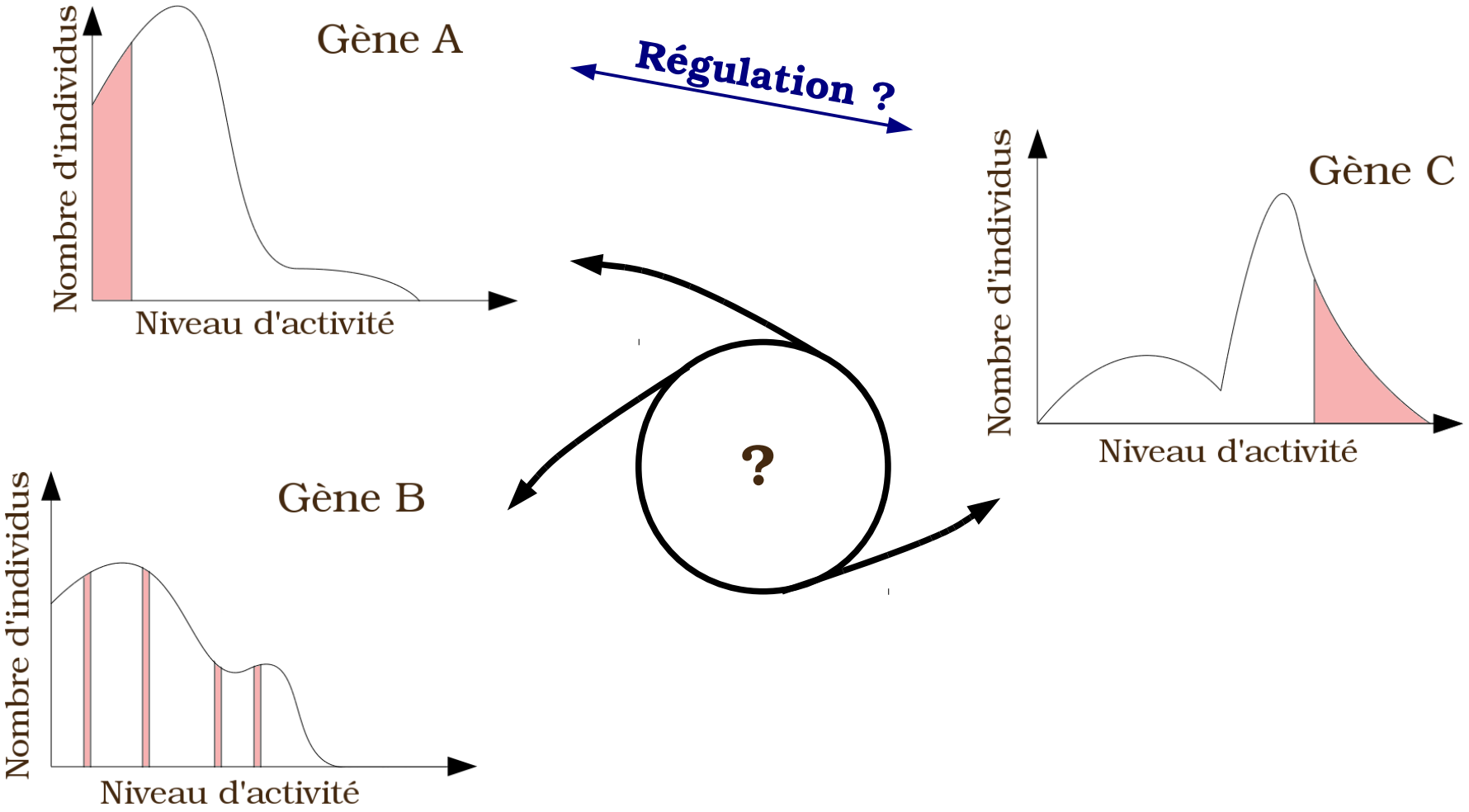
Motivation biologique

- Mesure pour un ensemble d'individus, l'activité de différents gènes (par exemple le niveau d'expression)



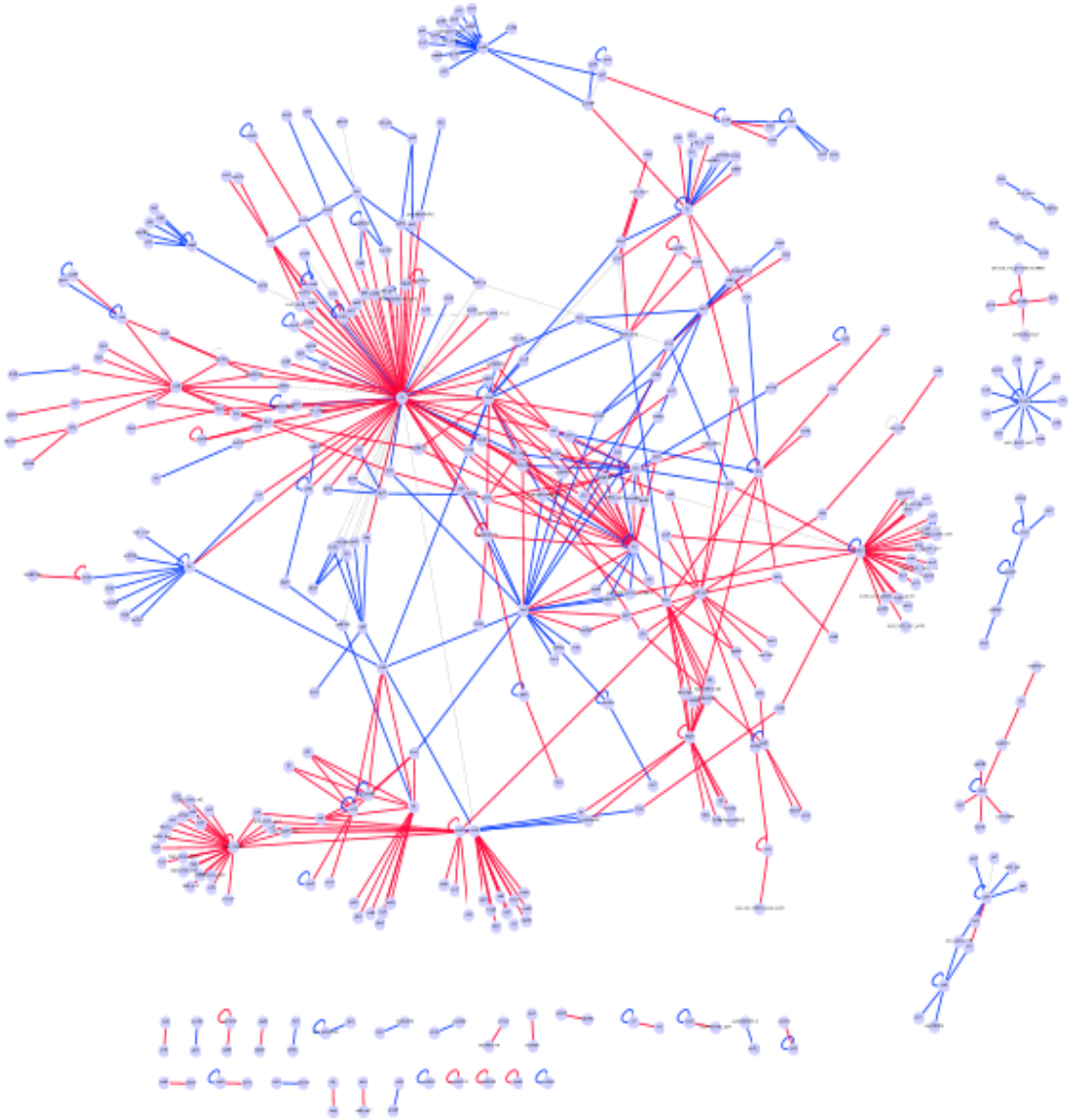
Motivation biologique

- Mesure pour un ensemble d'individus, l'activité de différents gènes (par exemple le niveau d'expression)



Objectif

- Apprendre la structure d'un réseau de régulation de gènes (RRG)



Objectif

- Apprendre la structure d'un réseau de régulation de gènes (RRG)
- Les réseaux bayésiens discrets, une solution ?

Objectif

- Apprendre la structure d'un réseau de régulation de gènes (RRG)
- Les réseaux bayésiens discrets, une solution ?

  - Modélisation de dépendances complexes (non-linéaires)
  - Estimation aisée des paramètres pour des données complètes
  - Modèle décomposable localement
  - Prise en compte de la causalité entre certaines variables

Objectif

- Apprendre la structure d'un réseau de régulation de gènes (RRG)
- Les réseaux bayésiens discrets, une solution ?

- Modélisation de dépendances complexes (non-linéaires)
- Estimation aisée des paramètres pour des données complètes
- Modèle décomposable localement
- Prise en compte de la causalité entre certaines variables

- Causalité partielle due aux équivalents de Markov
- Données d'expression continues → discrétisation nécessaire
- Ne peut représenter des circuits → approches ensemblistes
- Complexité pour apprendre la structure

Plan de l'exposé :

1 – Apprentissage de la structure d'un réseau bayésien

- Présentation des réseaux bayésiens
- État de l'art
- Nos propositions de nouveaux opérateurs locaux

2 – Application à la génétique-génomique

- Introduction aux réseaux de régulation de gènes
- État de l'art
- Notre modélisation des données de génétique-génomique
- Application aux données d'Arabidopsis *thaliana*

Plan de l'exposé :

1 – Apprentissage de la structure d'un réseau bayésien

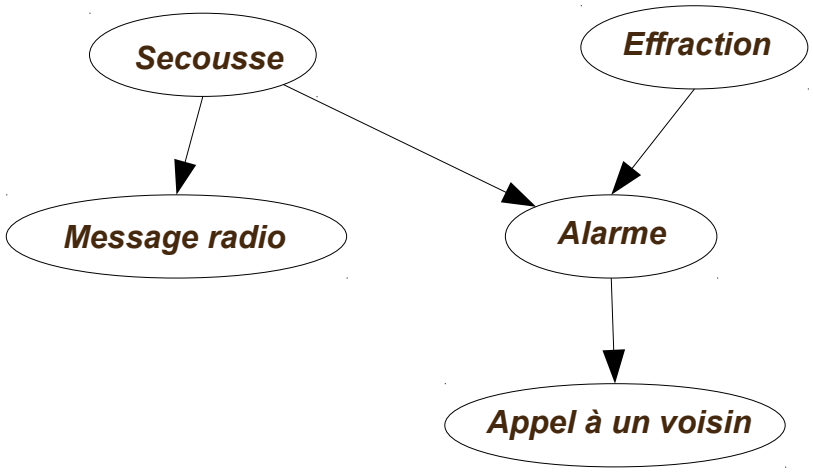
- **Présentation des réseaux bayésiens**
- **État de l'art**
- **Nos propositions de nouveaux opérateurs locaux**

2 – Application à la génétique-génomique

- Introduction aux réseaux de régulation de gènes
- État de l'art
- Notre modélisation des données de génétique-génomique
- Application aux données d'Arabidopsis *thaliana*

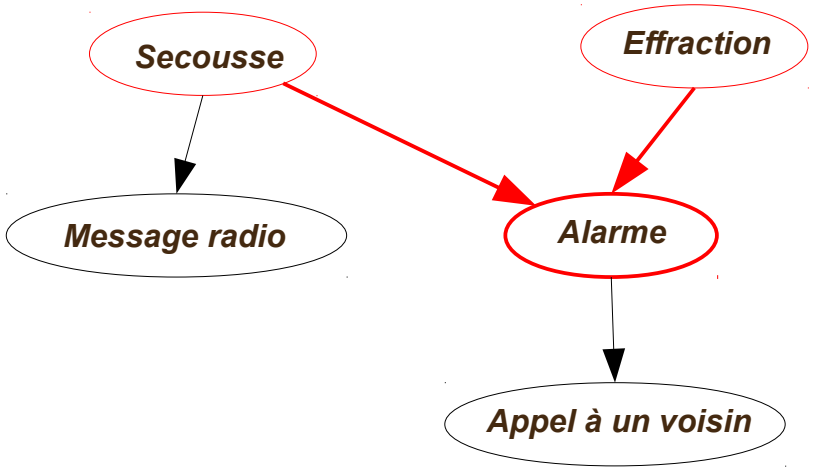
Les réseaux bayésiens

- Modèles graphiques probabilistes
  - **G**raphe **D**irigé **A**cyclique (DAG)
  - Probabilités conditionnelles d'un ensemble de p variables aléatoires X



Les réseaux bayésiens

- Modèles graphiques probabilistes
  - **G**raphe **D**irigé **A**cyclique (DAG)
  - Probabilités conditionnelles d'un ensemble de p variables aléatoires X



Secousse (S)	Effraction (E)	$P(Alarme/S,E)$	$P(!Alarme/S,E)$
Oui	Oui	0.90	0.10
Oui	Non	0.20	0.80
Non	Oui	0.90	0.10
Non	Non	0.01	0.99

- Pour chaque variable X_i on a $P_G(X_i/Pa_i^j)=\theta_i^j$

  où Pa_i : parents de X_i

- Distribution de probabilité jointe $P_G(X)=\prod_{i=1}^{p} P_G(X_i/Pa_i)$

Apprentissage de la structure

1 - Recherche des indépendances

- Rechercher les indépendances conditionnelles à l'aide d'un test statistique
  - Test du chi-2
  - Rapport de vraisemblance
  - Information mutuelle

  → nombre exponentiel de tests possibles
  → sensibilité à la puissance du test, risque de propagation des erreurs

2 - Méthodes à base de score

- Maximiser un score qui calcule la vraisemblance des données
  - Score BIC
  - Score BDe
  - Score fNML

  } *Scores équivalents,* } *pénalisés et décomposables*

  → problème d'optimisation NP-dur (Chickering et al., 02)

- Algorithmes de recherche locale

Recherches locales comparées

- *Voisinage d'un graphe G* : l'ensemble des graphes atteignables à partir de G par l'application d'opérateurs élémentaires
- *Opérateurs élémentaires* : - ajout d'un arc
  - suppression d'un arc
  - inversion d'un arc

Recherches locales comparées

- *Voisinage d'un graphe G* : l'ensemble des graphes atteignables à partir de G par l'application d'opérateurs élémentaires
- *Opérateurs élémentaires* : - ajout d'un arc
  - suppression d'un arc
  - inversion d'un arc

- Stratégie de déplacement dans le voisinage
  - Approches gloutonnes
    - → Concept algorithmique simple
    - → Peu de paramètres
    - → Bonnes performances en pratique

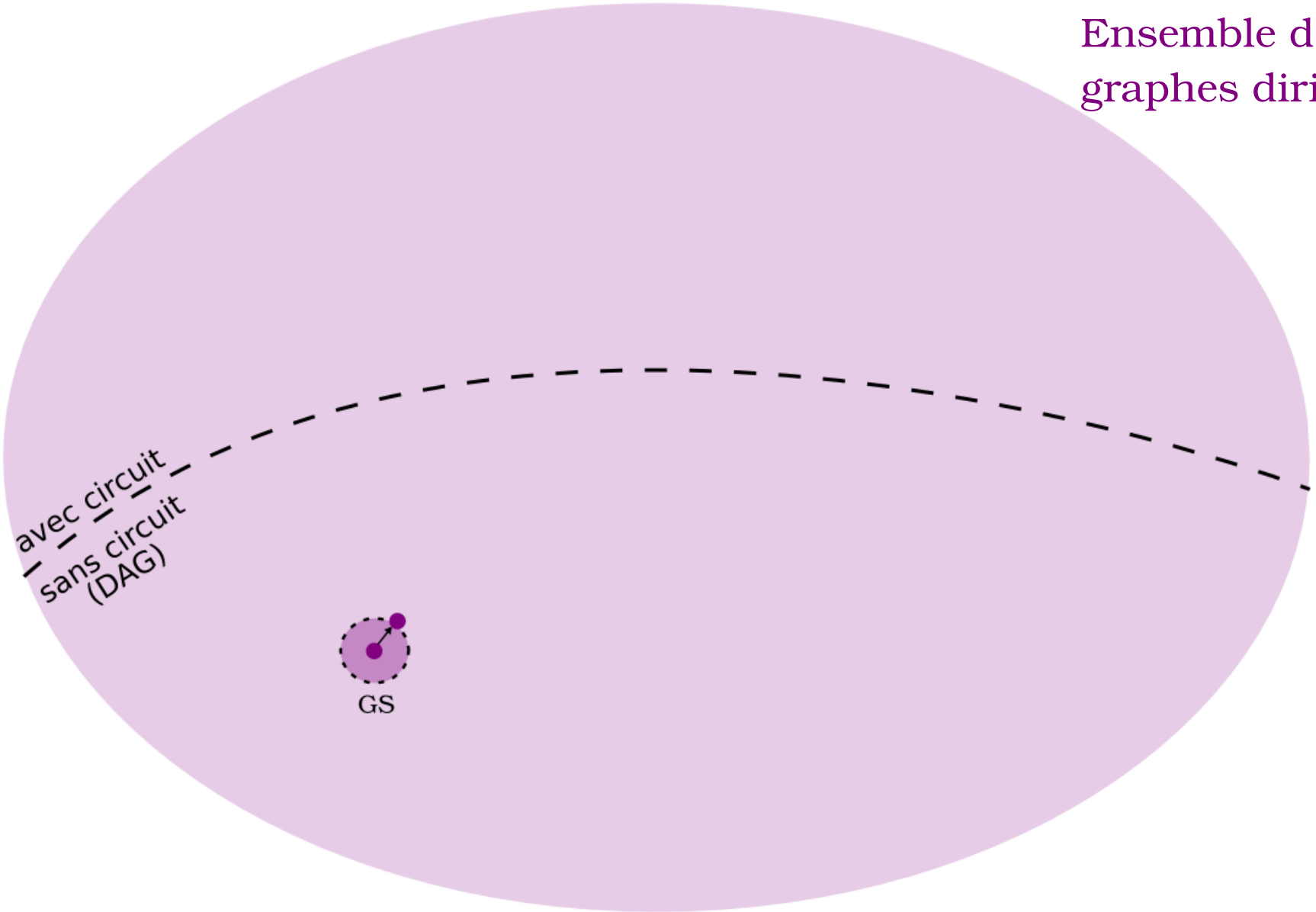
Recherches locales comparées

- *Voisinage d'un graphe G* : l'ensemble des graphes atteignables à partir de *G* par l'application d'opérateurs élémentaires
- *Opérateurs élémentaires* : - ajout d'un arc
  - suppression d'un arc
  - inversion d'un arc

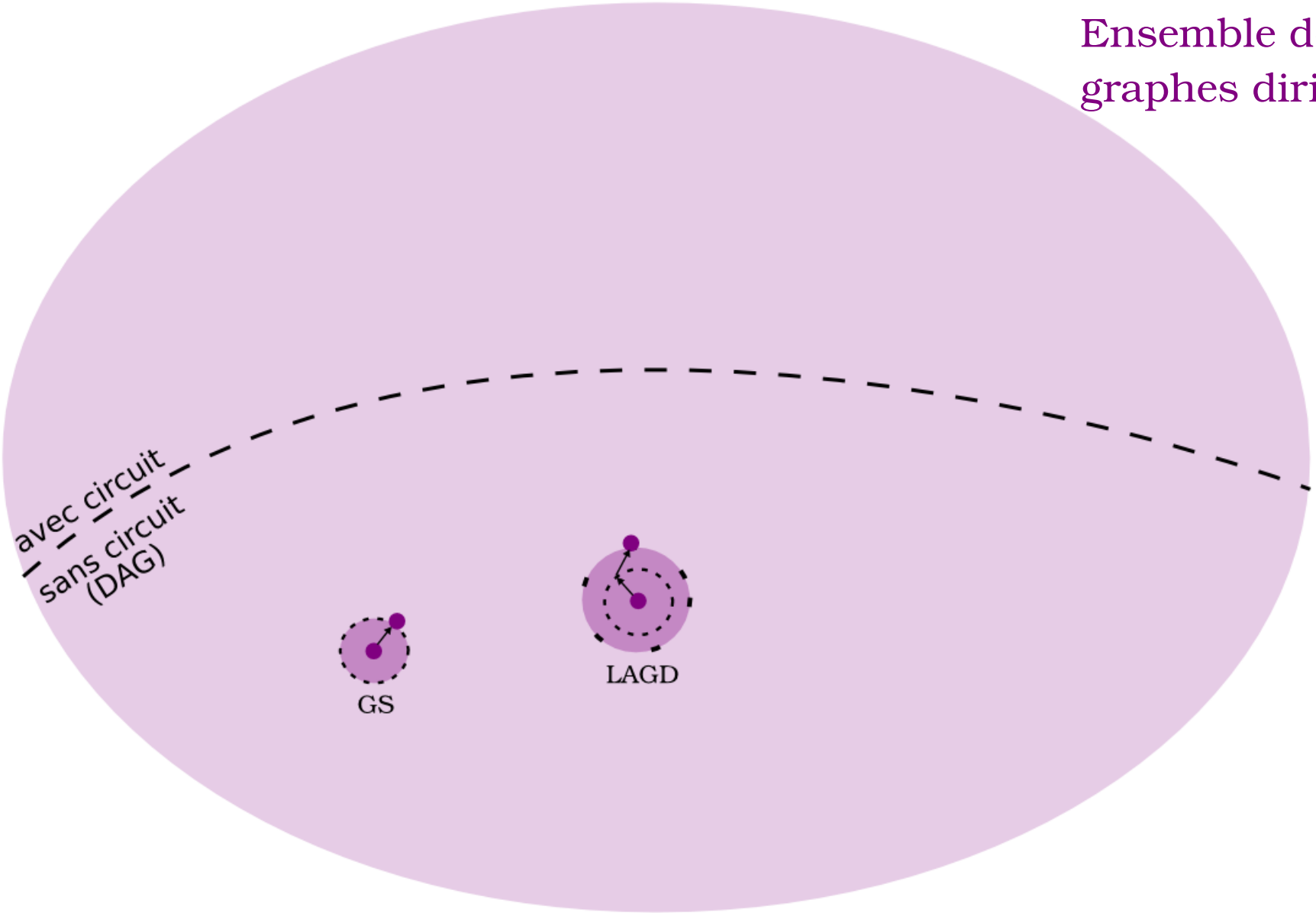
- Stratégie de déplacement dans le voisinage
  - Approches gloutonnes
    - → Concept algorithmique simple
    - → Peu de paramètres
    - → Bonnes performances en pratique

- Espace de recherche
  - Espace des DAG
    - → GS, **notre contribution SGS**
    - → LAGD (Holland et al., 08)
  - Espace des équivalents de Markov
    - → GES (Chickering et al., 02)

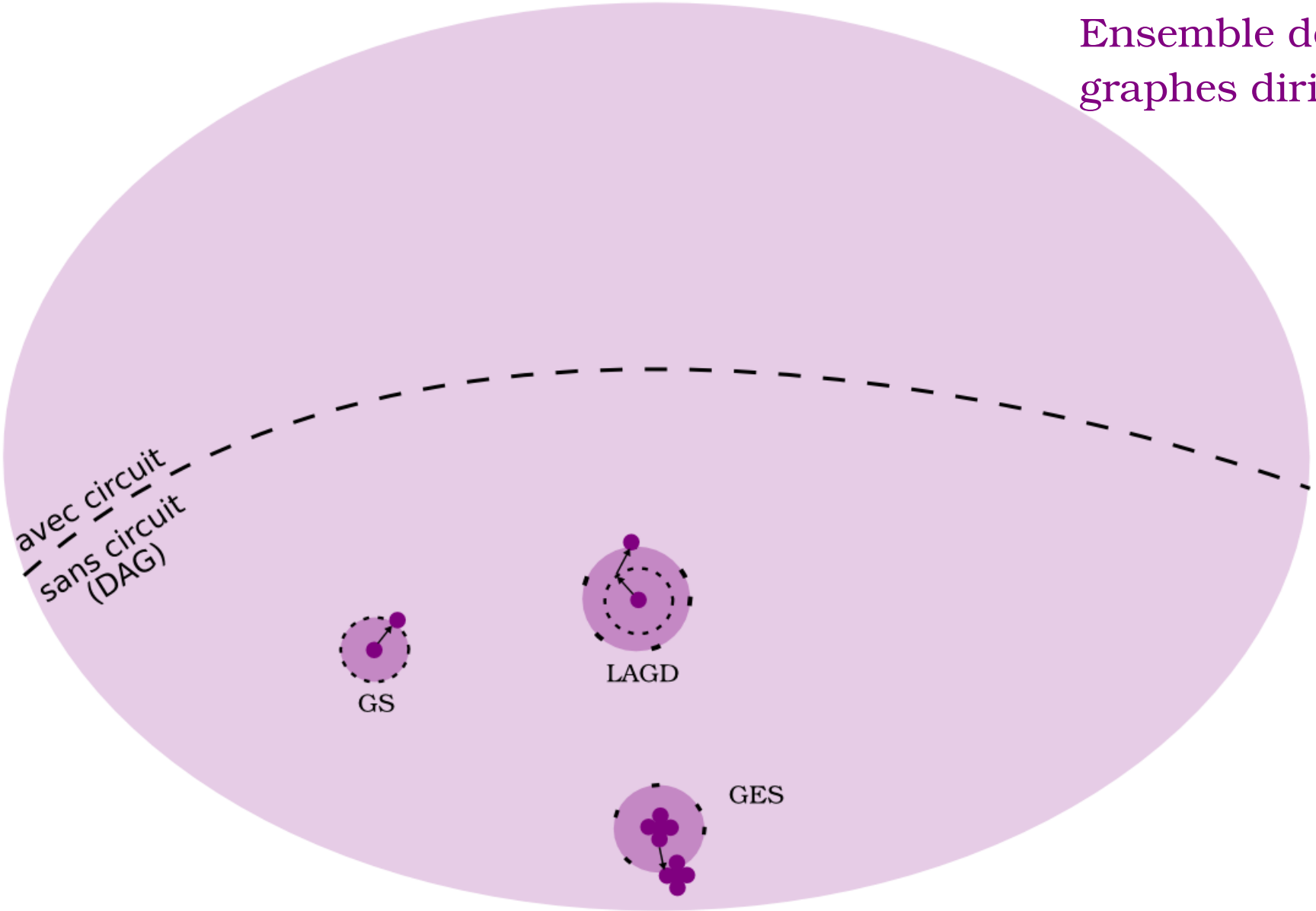
Espace de recherche



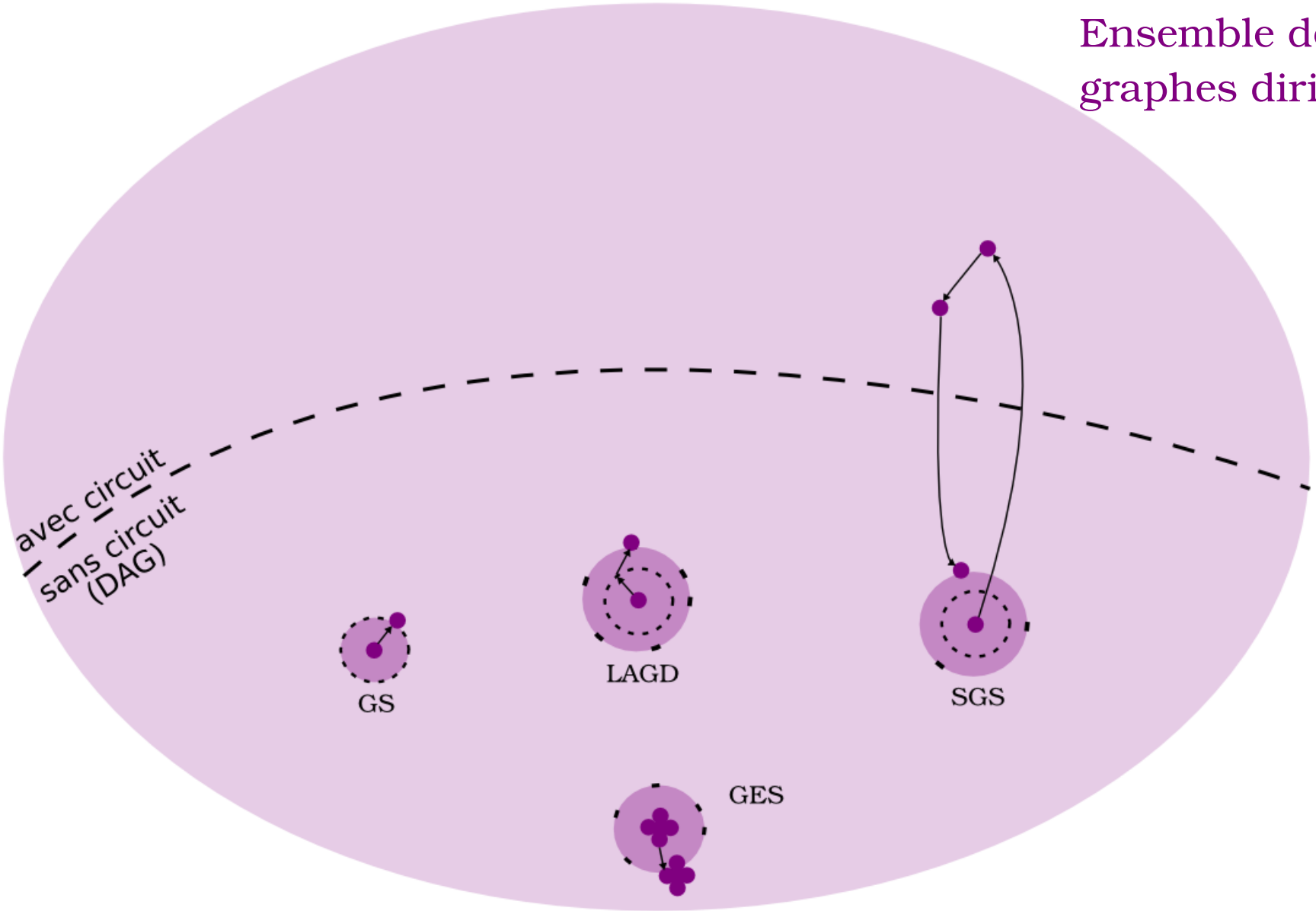
Espace de recherche



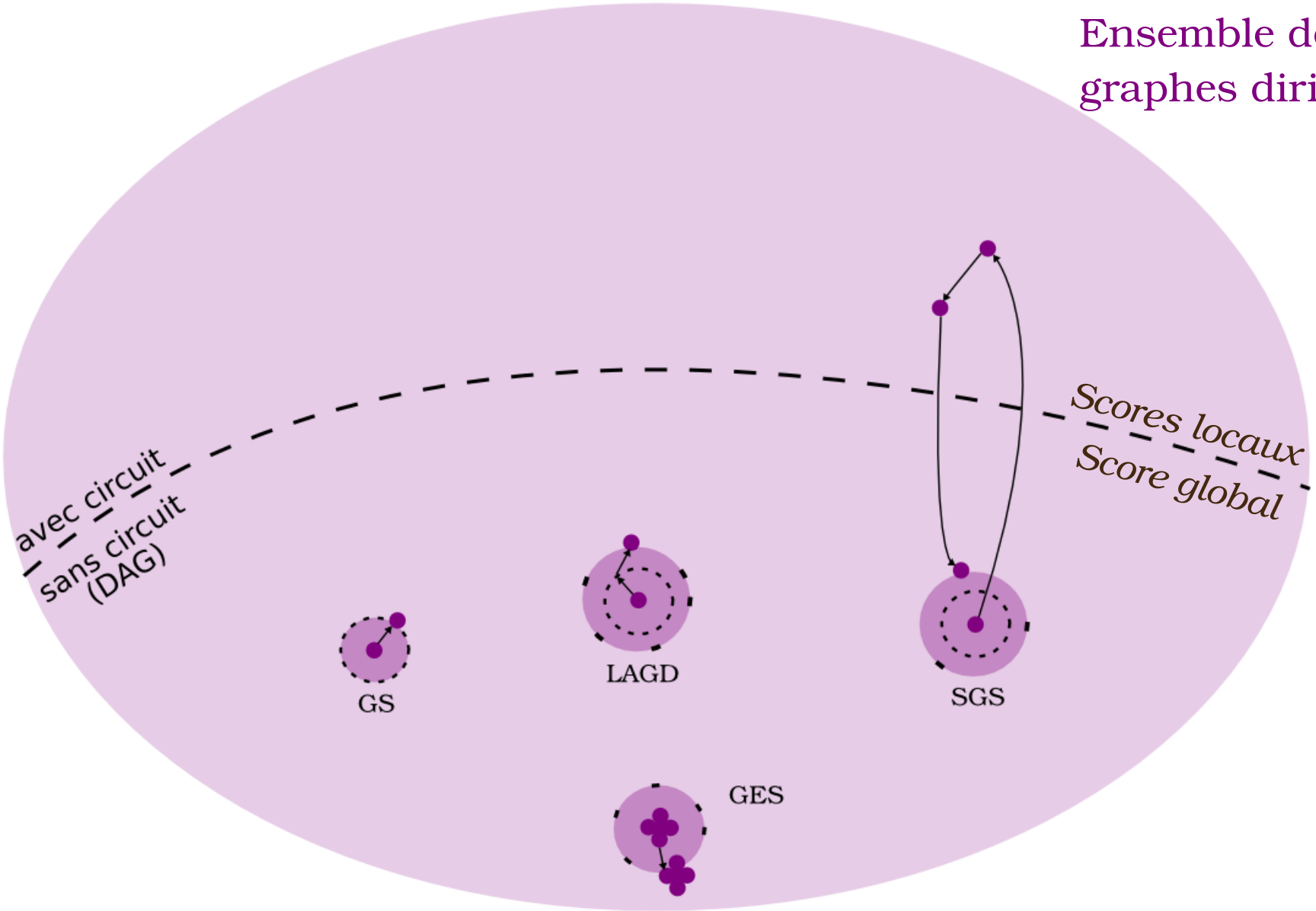
Espace de recherche



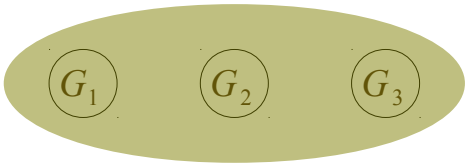
Espace de recherche



Espace de recherche



Greedy Search (GS)

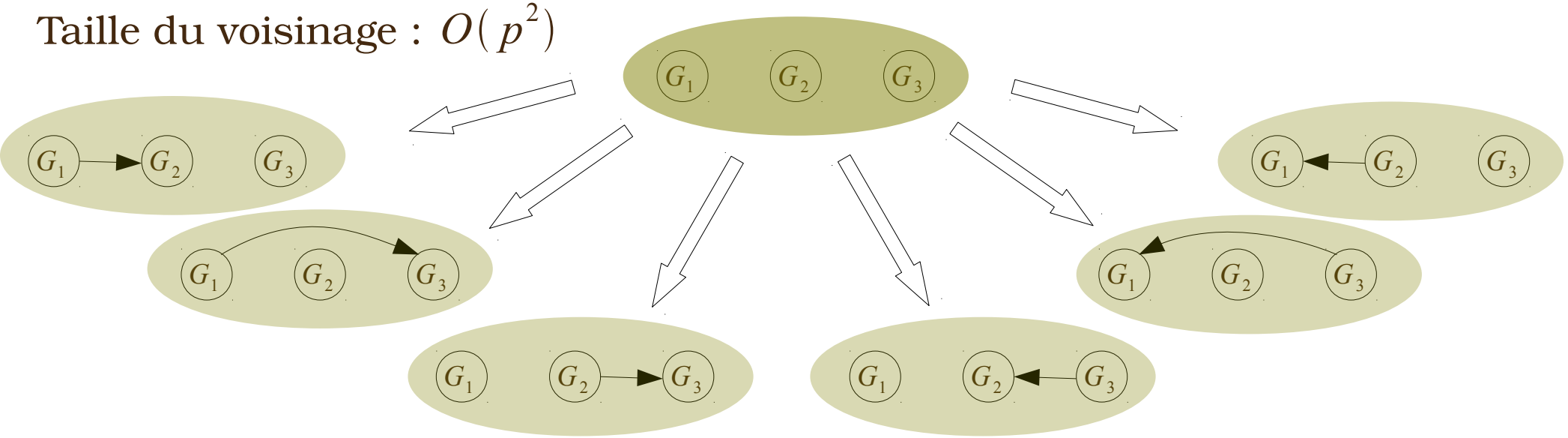


3 opérateurs classiques

- Ajout d'un arc
- Suppression d'un arc
- Inversion d'un arc

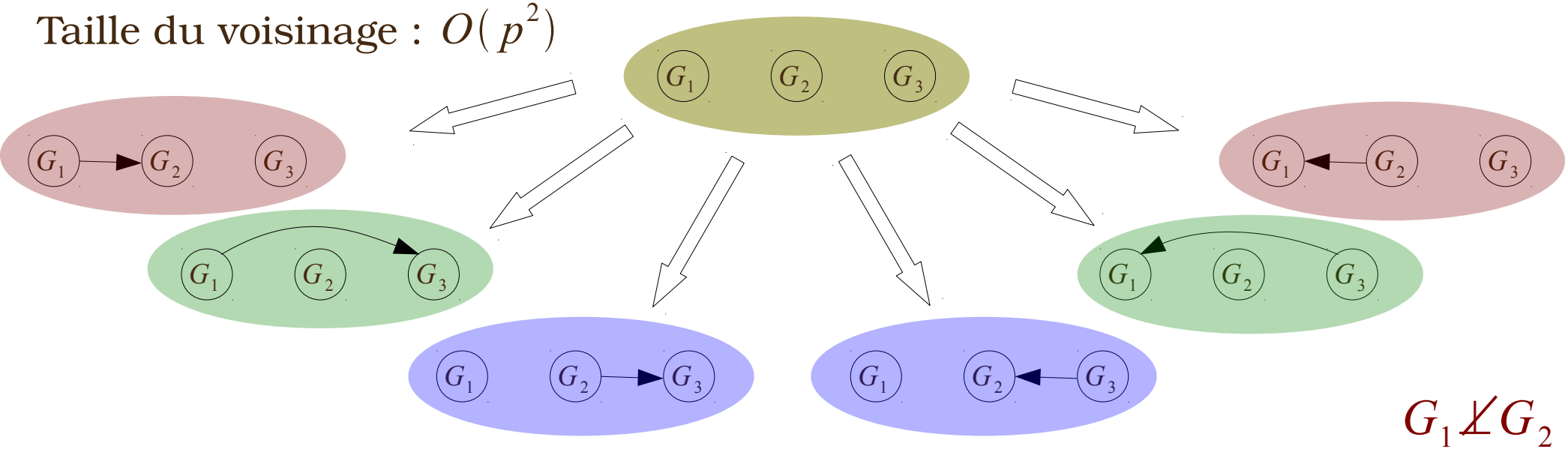
Greedy Search (GS)

Taille du voisinage : $O(p^2)$



Greedy Search (GS)

Taille du voisinage : $O(p^2)$



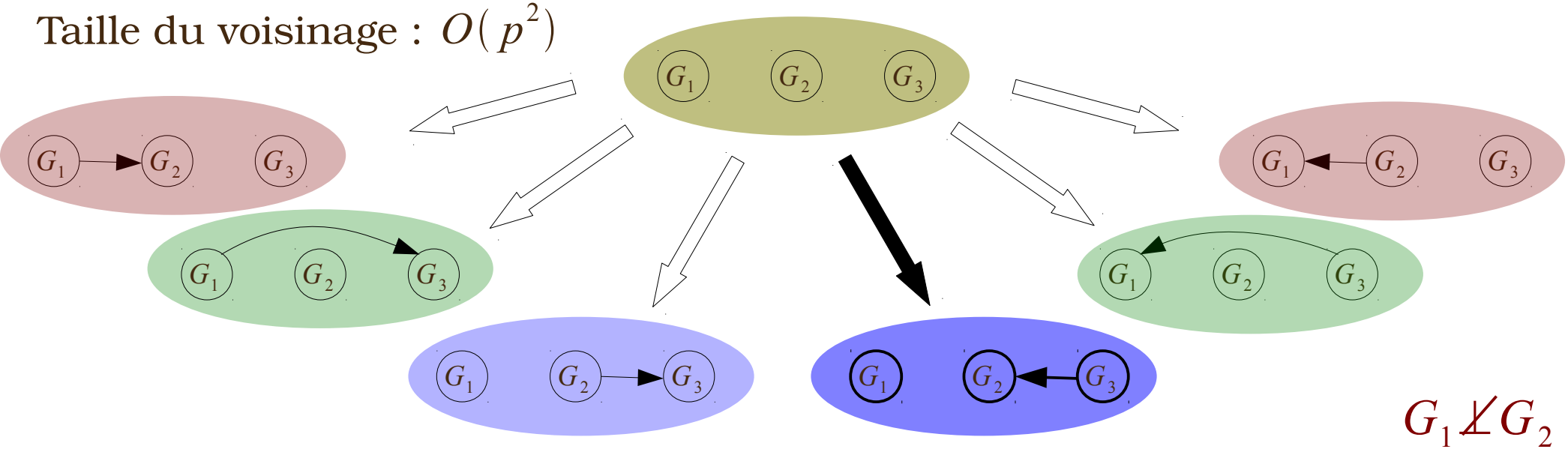
$G_1 \not\perp G_2$

$G_1 \not\perp G_3$

$G_2 \not\perp G_3$

Greedy Search (GS)

Taille du voisinage : $O(p^2)$



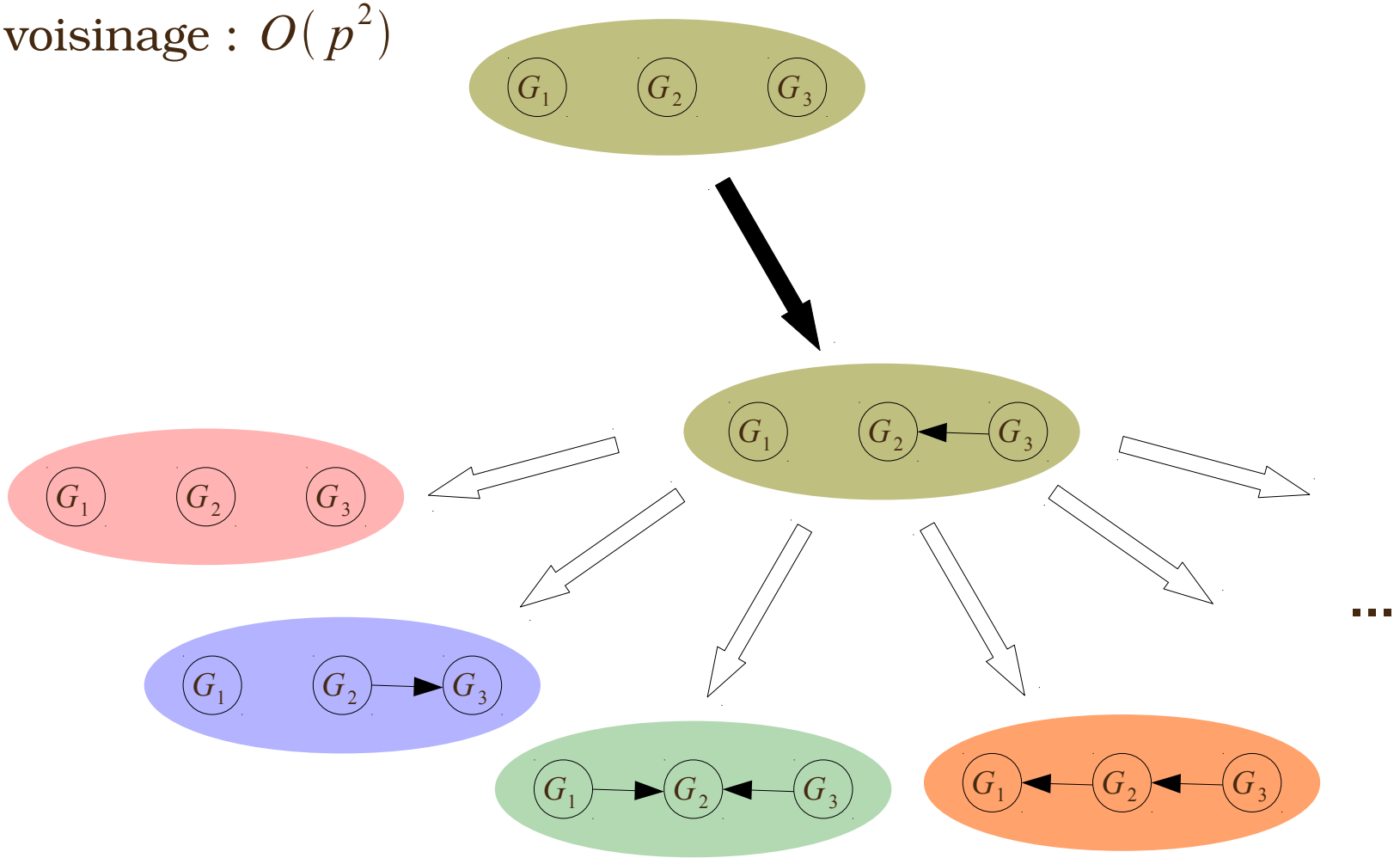
$G_1 \not\perp G_2$

$G_1 \not\perp G_3$

$G_2 \not\perp G_3$

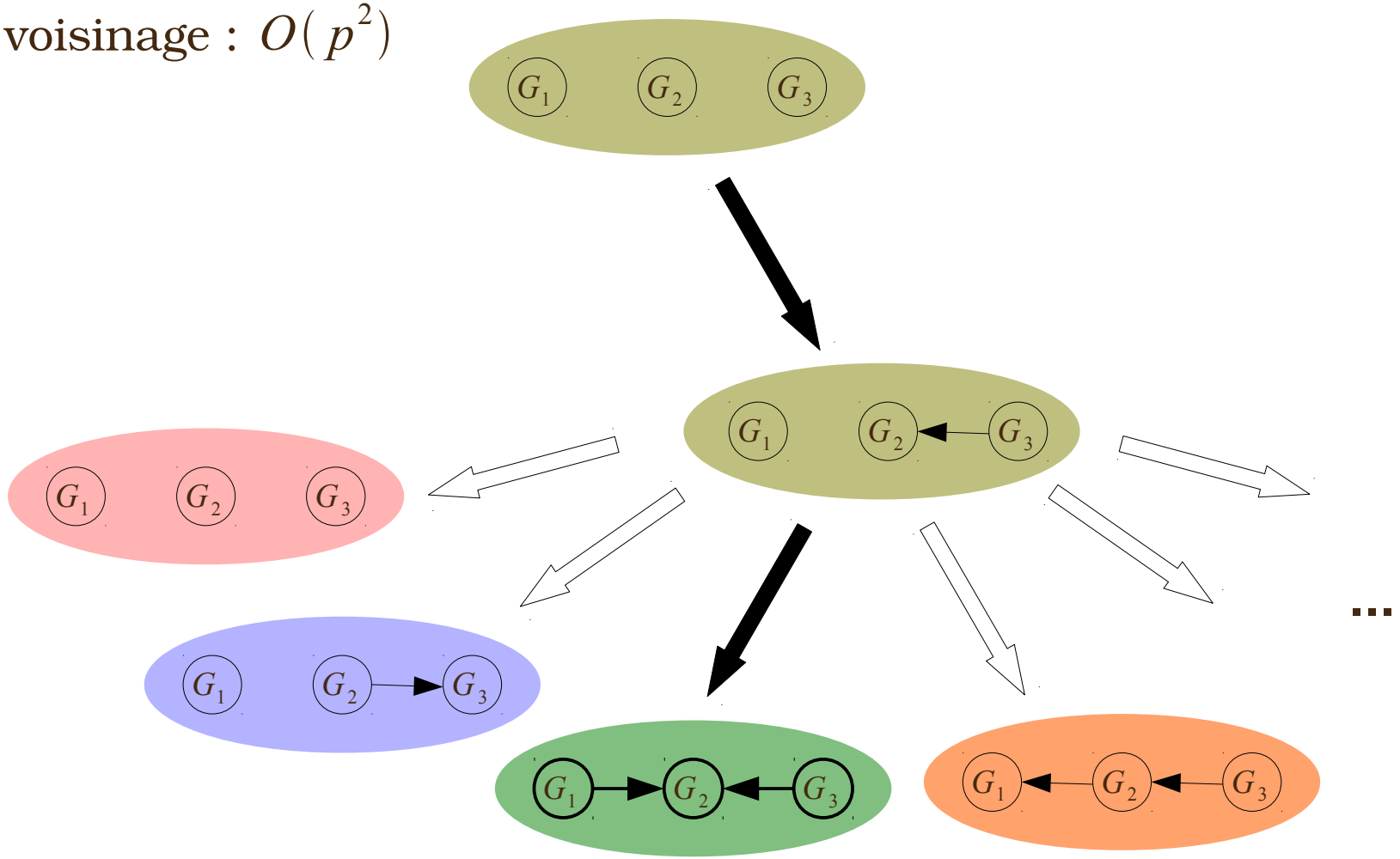
Greedy Search (GS)

Taille du voisinage : $O(p^2)$



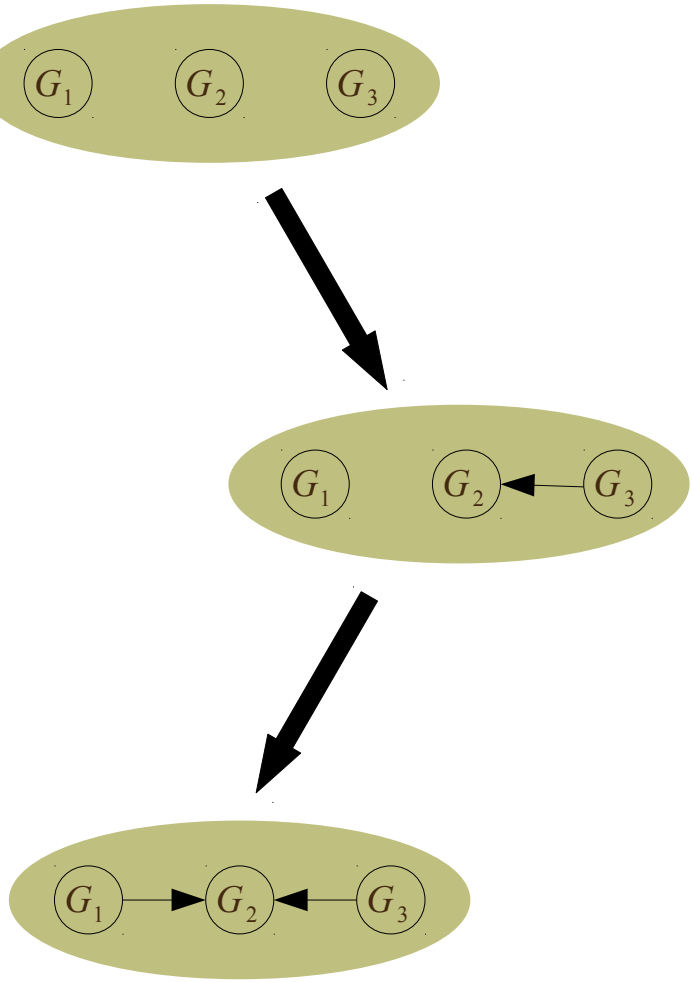
Greedy Search (GS)

Taille du voisinage : $O(p^2)$



Greedy Search (GS)

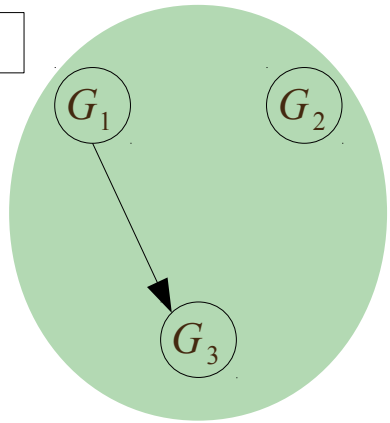
Taille du voisinage : $O(p^2)$



Opérateur SWAP

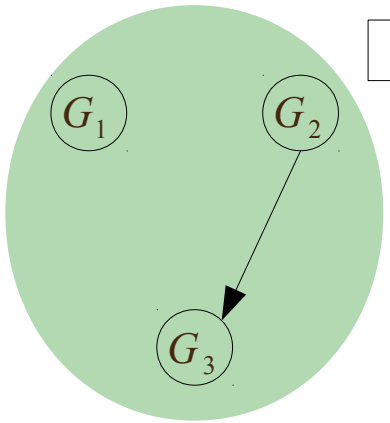
Graphe courant

17



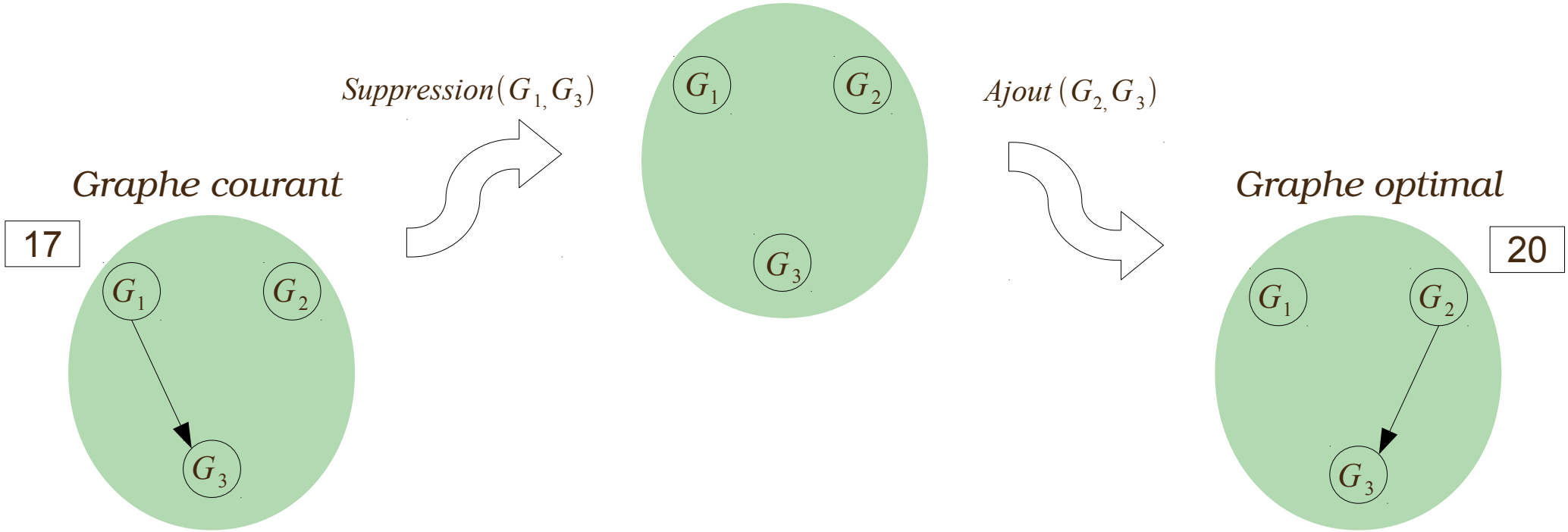
Graphe optimal

20



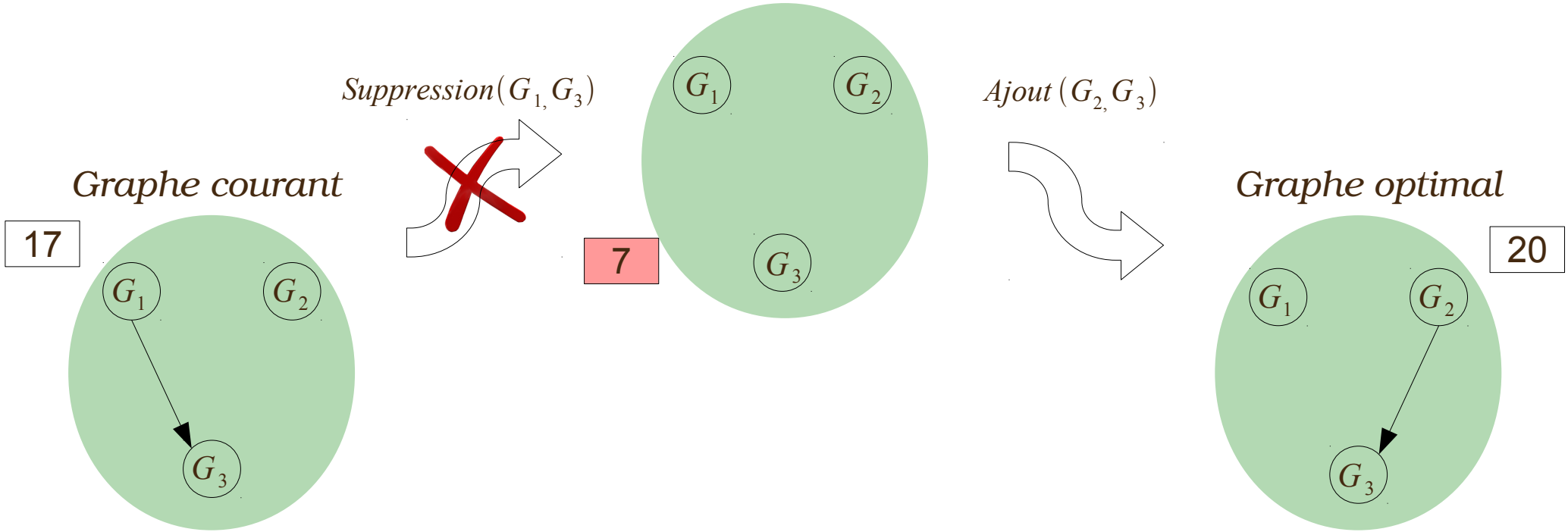
→ échapper aux solutions de maximum local

Opérateur SWAP



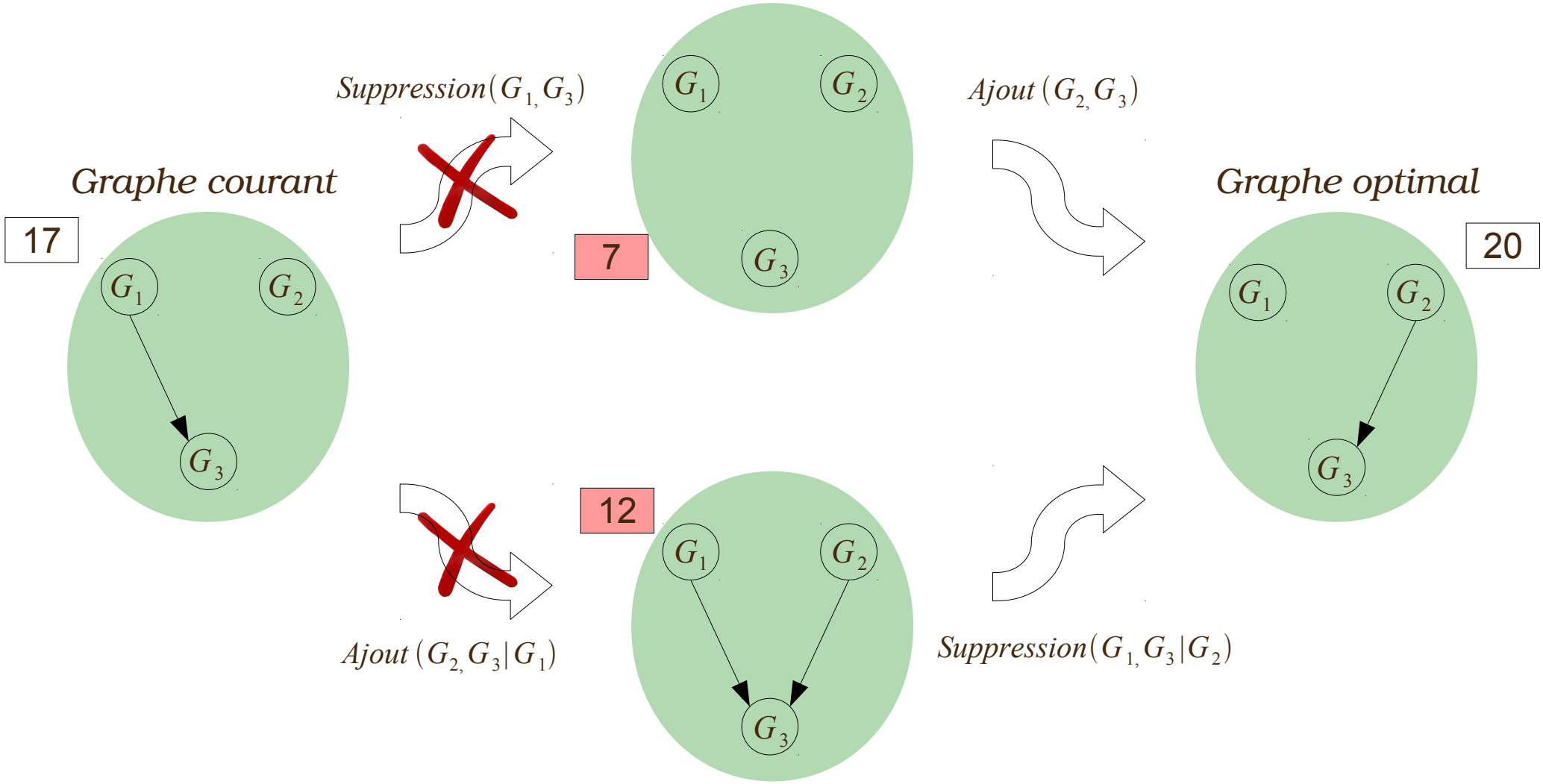
→ échapper aux solutions de maximum local

Opérateur SWAP



→ échapper aux solutions de maximum local

Opérateur SWAP



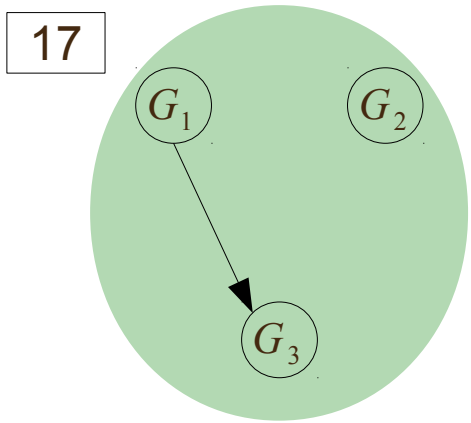
→ échapper aux solutions de maximum local

Opérateur SWAP

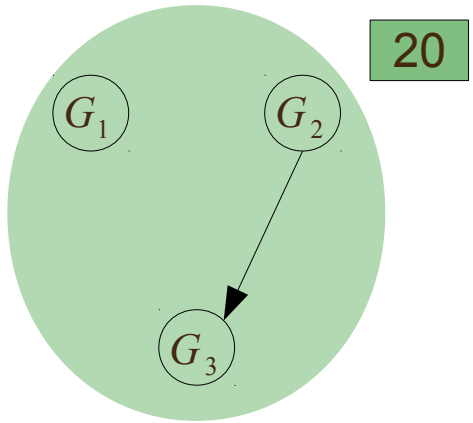
Taille du voisinage : $O(kp^2)$

avec k le nombre maximum de parents

Graphe courant



Graphe optimal

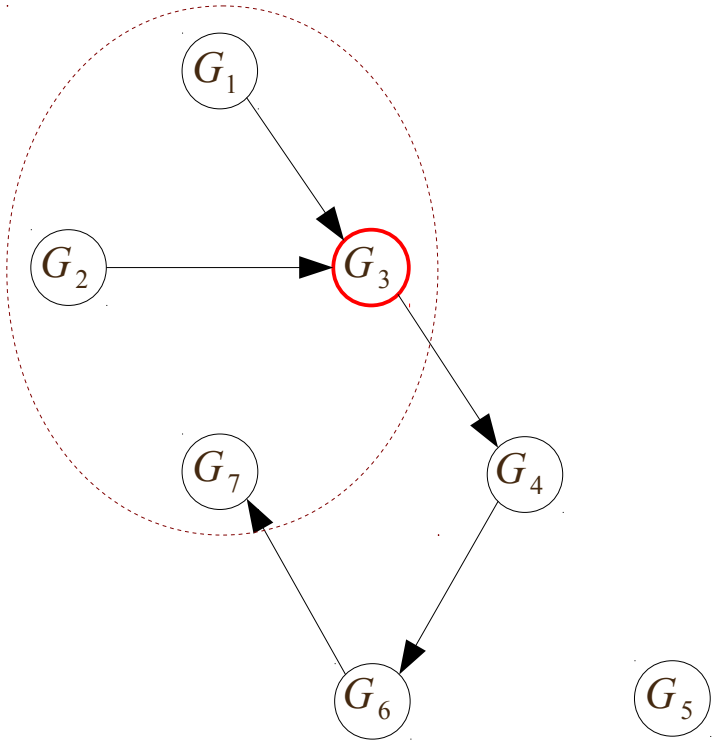


→ échapper aux solutions de maximum local

Opérateur Itératif

$Swap(G_2, \boldsymbol{G_3}, G_7)?$

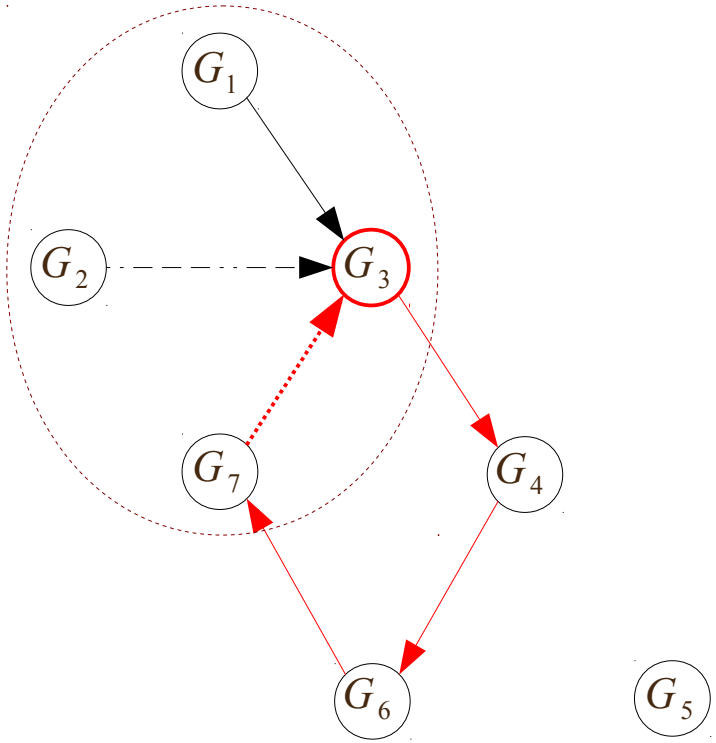
Graphe courant



Opérateur Itératif

$$Swap(G_2, \mathbf{G_3}, G_7)? \longrightarrow Circuit\{G_3, G_4, G_6, G_7\}$$

Graphe courant

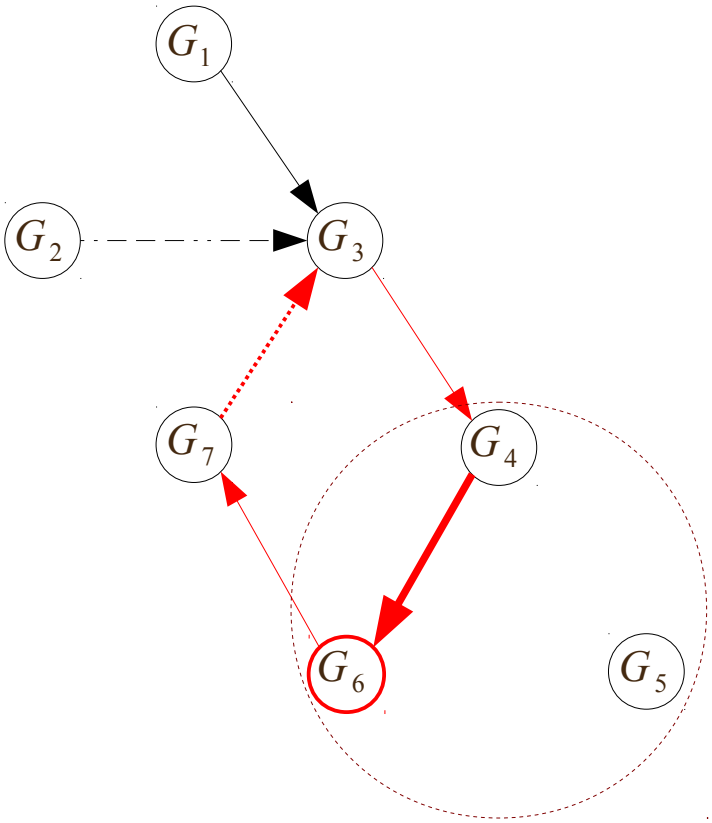


Objectif : supprimer les circuits

Opérateur Itératif

$$Swap(G_2, \mathbf{G_3}, G_7)? \longrightarrow Circuit\{G_3, G_4, G_6, G_7\}$$

Graphe courant



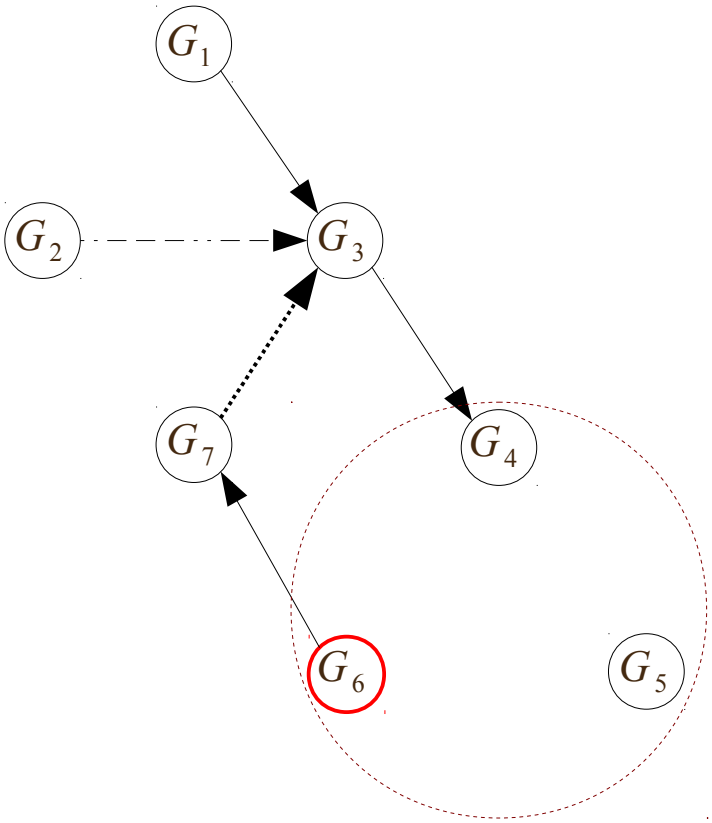
Objectif : supprimer les circuits

Étape 1. Supprimer l'arc qui minimise le score local

Opérateur Itératif

$$Swap(G_2, \mathbf{G_3}, G_7)? \longrightarrow Circuit\{G_3, G_4, G_6, G_7\}$$

Graphe courant



Objectif : supprimer les circuits

Étape 1. Supprimer l'arc qui minimise le score local

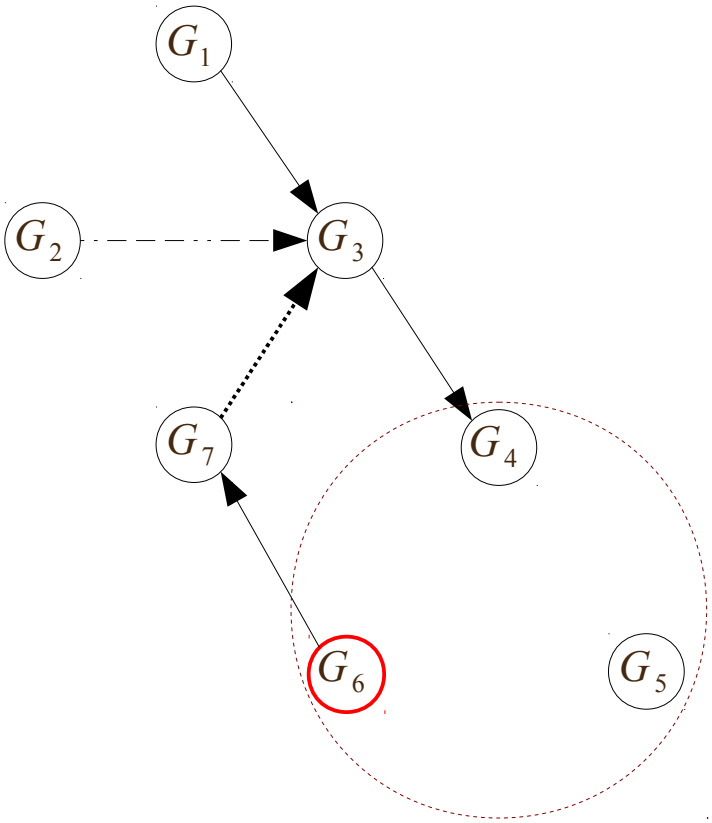
Amélioration du score ?

Oui

Opérateur Itératif

$$Swap(G_2, \mathbf{G_3}, G_7)? \longrightarrow Circuit\{G_3, G_4, G_6, G_7\}$$

Graphe courant



Objectif : supprimer les circuits

Étape 1. Supprimer l'arc qui minimise le score local

Amélioration du score ?

Oui → **Aucun circuit ?**

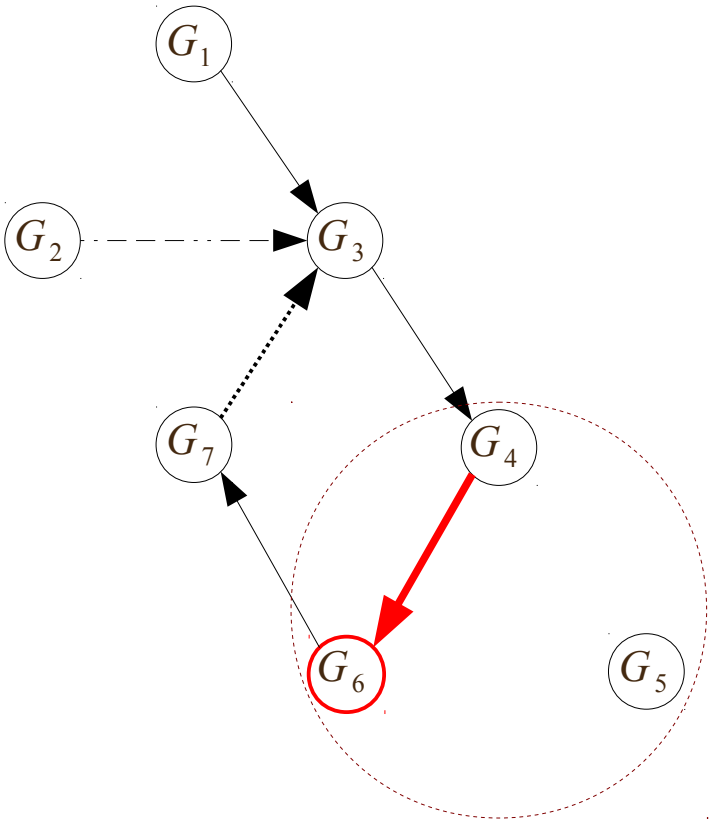
Oui → OK !

Non → Retour à l'étape1

Opérateur Itératif

$$Swap(G_2, \mathbf{G_3}, G_7)? \longrightarrow Circuit\{G_3, G_4, G_6, G_7\}$$

Graphe courant



Objectif : supprimer les circuits

Étape 1. Supprimer l'arc qui minimise le score local

Amélioration du score ?

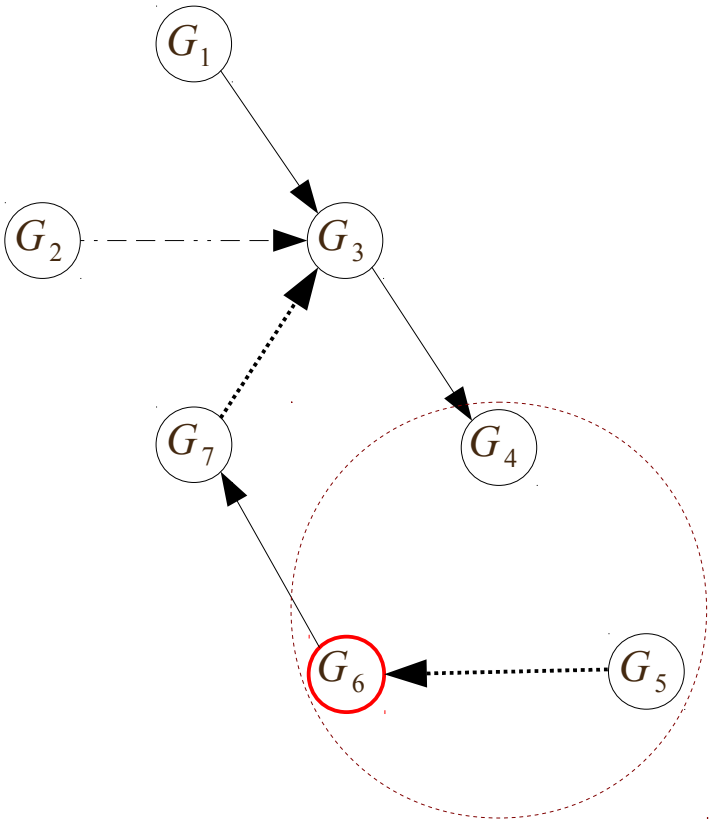
- Oui → Aucun circuit ?
  - Oui → OK !
  - Non → Retour à l'étape1
- **Non → Continuer l'étape 2**

Étape 2. *Swapper* cet arc

Opérateur Itératif

$$Swap(G_2, \mathbf{G_3}, G_7)? \longrightarrow Circuit\{G_3, G_4, G_6, G_7\}$$

Graphe courant



Objectif : supprimer les circuits

Étape 1. Supprimer l'arc qui minimise le score local

Amélioration du score ?

- Oui → Aucun circuit ?
  - Oui → OK !
  - Non → Retour à l'étape1
- Non → Continuer l'étape 2

Étape 2. *Swapper* cet arc

Amélioration du score ?

Oui → Aucun circuit ?

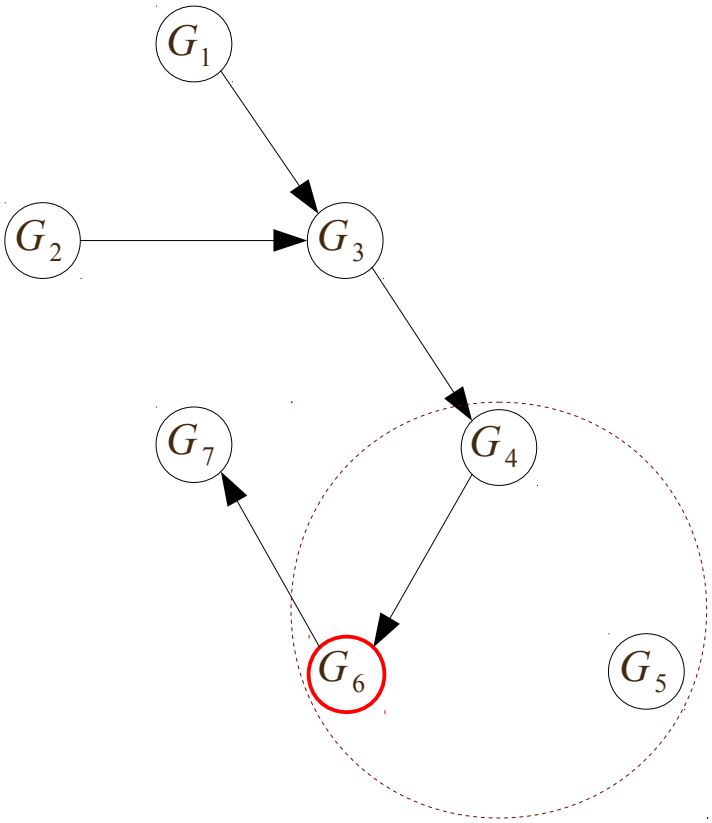
Oui → OK !

Non → Retour à l'étape 1

Opérateur Itératif

$$Swap(G_2, \mathbf{G_3}, G_7)? \longrightarrow Circuit\{G_3, G_4, G_6, G_7\}$$

Graphe courant



Objectif : supprimer les circuits

Étape 1. Supprimer l'arc qui minimise le score local

Amélioration du score ?

- Oui → Aucun circuit ?
  - Oui → OK !
  - Non → Retour à l'étape1
- Non → Continuer l'étape 2

Étape 2. *Swapper* cet arc

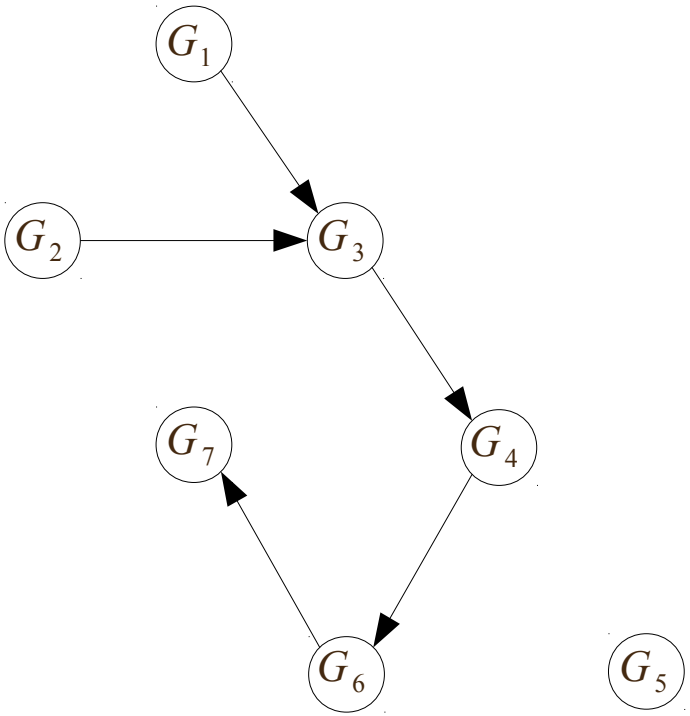
Amélioration du score ?

- Oui → Aucun circuit ?
  - Oui → OK !
  - Non → Retour à l'étape 1
- **Non → Game Over !**

Opérateur Itératif

$$Swap(G_2, \mathbf{G_3}, G_7)? \longrightarrow Circuit\{G_3, G_4, G_6, G_7\}$$

Graphe courant



Objectif : supprimer les circuits

Étape 1. **Supprimer** l'arc qui minimise le score local

Amélioration du score ?

- Oui → Aucun circuit ?
  - Oui → OK !
  - Non → Retour à l'étape1
- Non → Continuer l'étape 2

Étape 2. ***Swapper*** cet arc

Amélioration du score ?

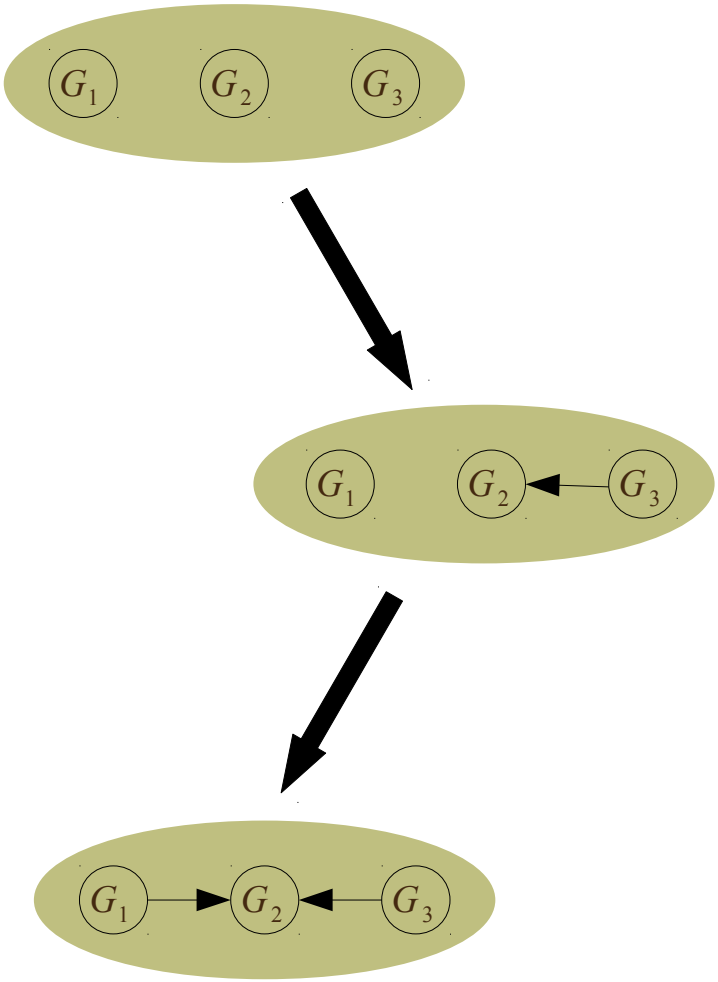
- Oui → Aucun circuit ?
  - Oui → OK !
  - Non → Retour à l'étape 1
- Non → Game Over !

Nombre borné d'opérations : $O(kp)$

avec k le nombre maximum de parents

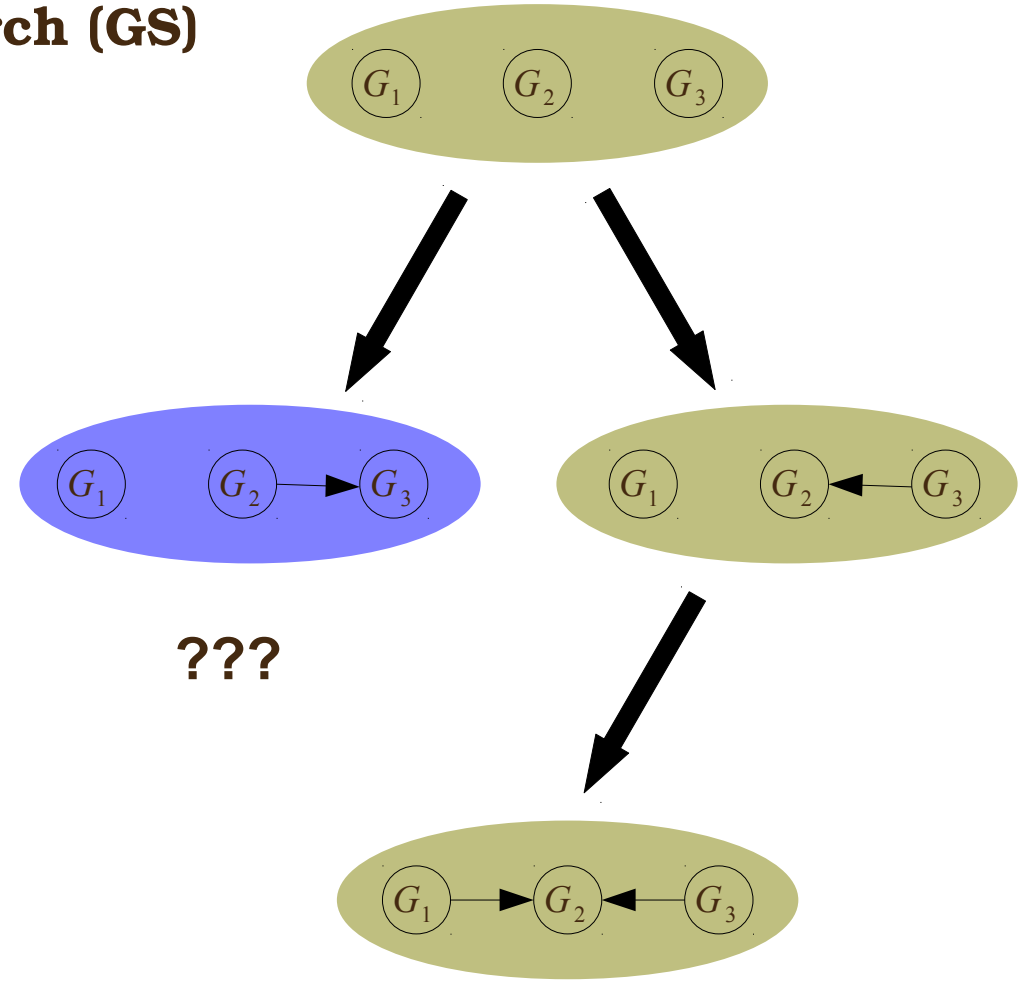
Stochastic Greedy Search

- **Greedy Search (GS)**



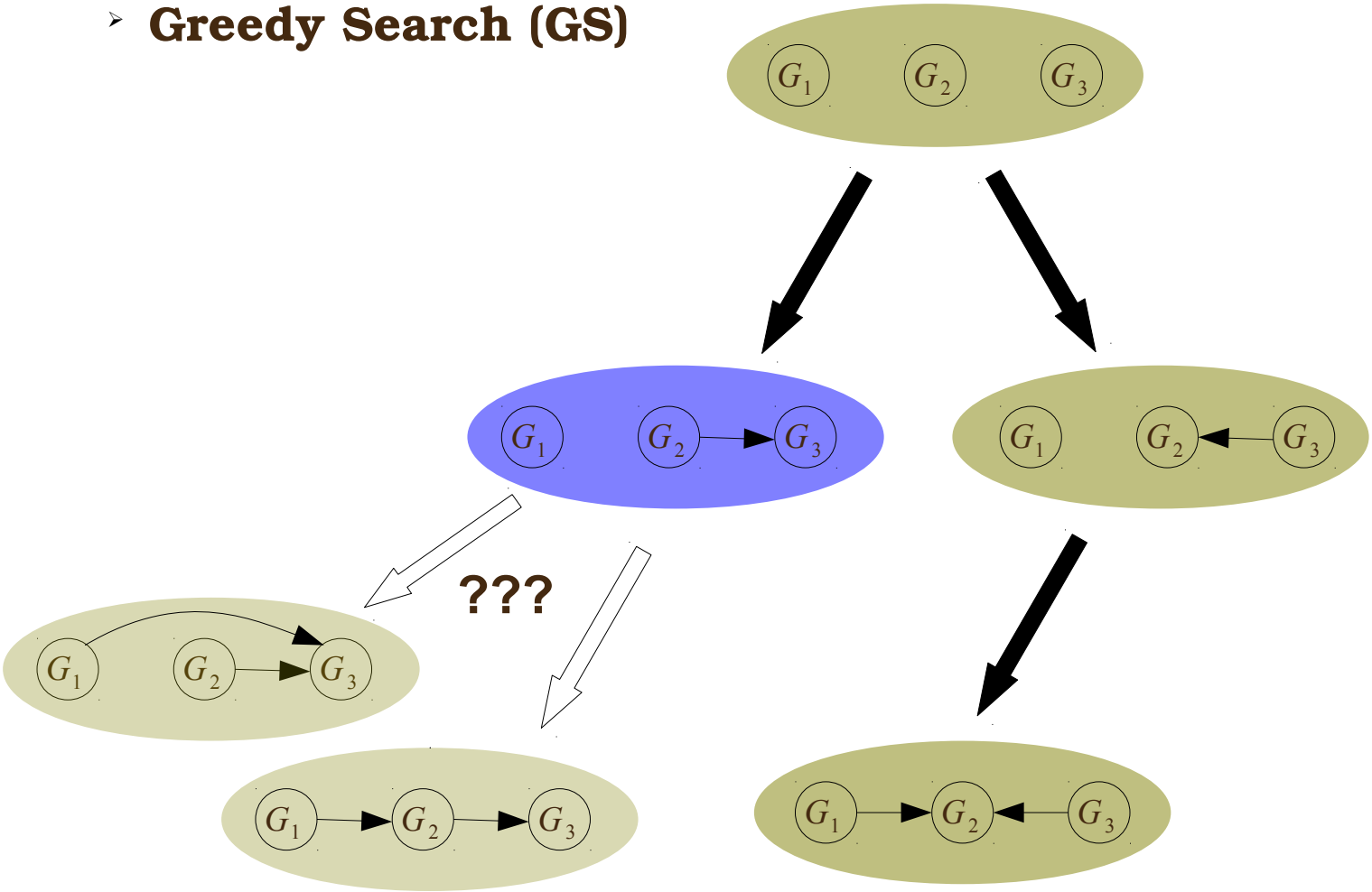
Stochastic Greedy Search

- **Greedy Search (GS)**



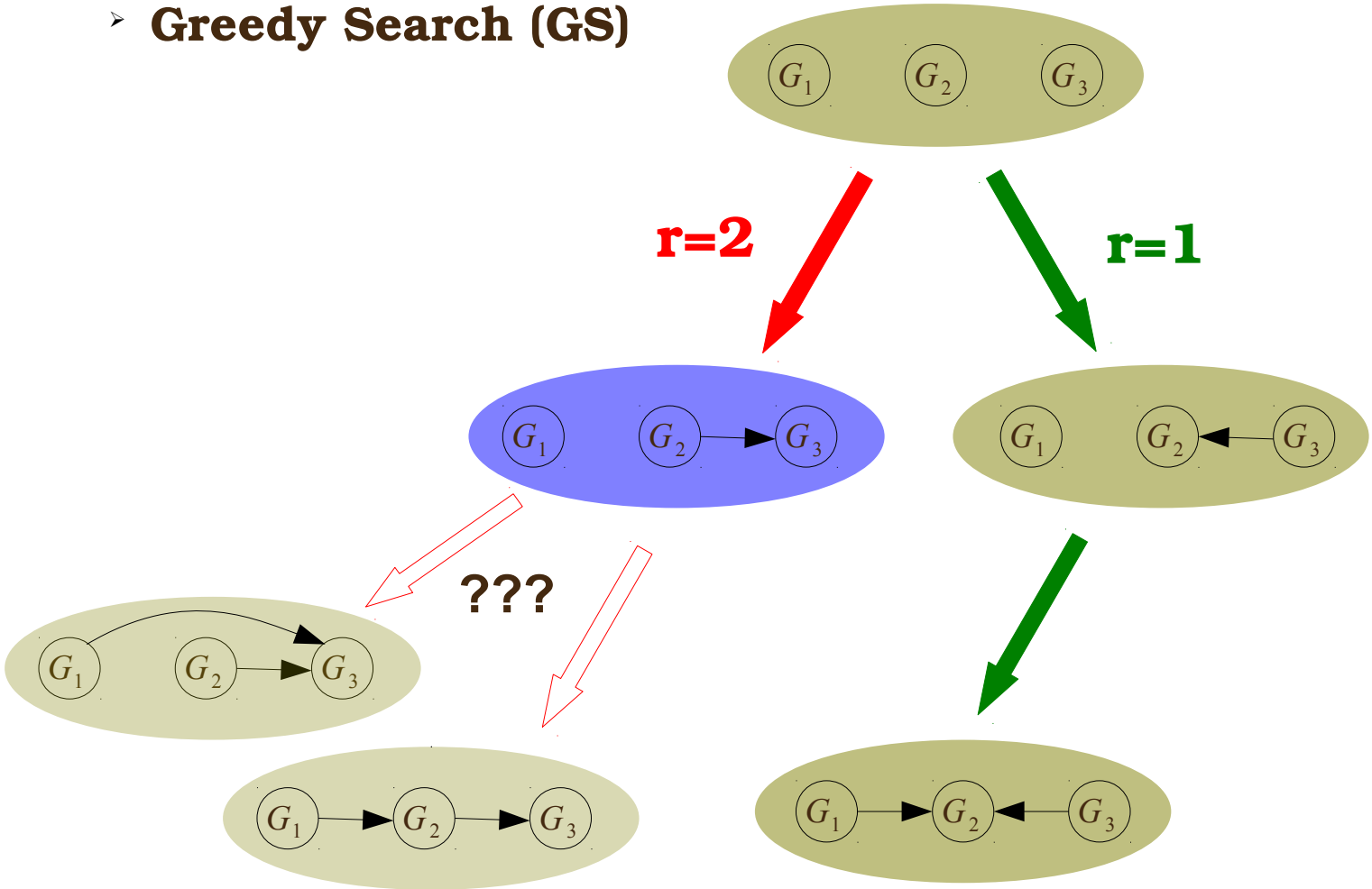
Stochastic Greedy Search

- **Greedy Search (GS)**



Stochastic Greedy Search

- **Greedy Search (GS)**



- **Stochastic Greedy Search (SGS)**

  = GS + orientation aléatoire des arcs inversibles

  → Parcours d'équivalents de Markov différents à chaque répétition (**r**)

Comparatifs

- 4 réseaux benchmarks

	Alarm	*Insurance*	*Hailfinder*	*Pigs*
Nb. de variables	37	27	56	441
Nb. d'arcs	46	52	66	592
Degré entrant	4	3	4	2

- Données générées à partir de distributions de probabilité :
  100 jeux de données (500 → 5 000 exemples)

- **SGS** (+SWAP+opérateurs itératifs) comparé à : - **LAGD**
  - **GES**

- Nombre limité de parents : 5

- Pré-sélection de parents potentiels pour le réseau Pigs avec SGS
  $\Delta\, Add(Parent, Cible) > 0$

Comparatifs : Scores atteints

- Comparaison des meilleurs scores BDeu atteints en moyenne sur les 100 jeux de données
- **SGS** et **LAGD** : meilleur score atteint sur 10 répétitions (r=10)

Test de Wilcoxon 5%	*Alarm*		*Insurance*		*Hailfinder*		*Pigs*	
	500	5 000	500	5 000	500	5 000	500	5 000
SGS vs **GES**	+	+	+	+	+	+	+	-
SGS vs **LAGD**	+	+	+	+	~	+	n/a	n/a
LAGD vs **GES**	+	~	+	+	+	+	n/a	n/a

Comparatifs : Distances d'édition

- Comparaison des structures non-orientées via les distances d'édition moyennes sur les 100 jeux de données
- **SGS** et **LAGD** : meilleur score atteint sur 10 répétitions (r=10)

	Alarm		*Insurance*		*Hailfinder*		*Pigs*	
	500	5 000	500	5 000	500	5 000	500	5 000
SGS	**11***	8	**24***	**10***	41	**29***	32	41
LAGD	15	10	**24***	16	47	39	n/a	n/a
GES	**11***	**6***	25	15	**39***	33	**9***	**0***

*** meilleur résultat**

Comparatifs : Distances d'édition

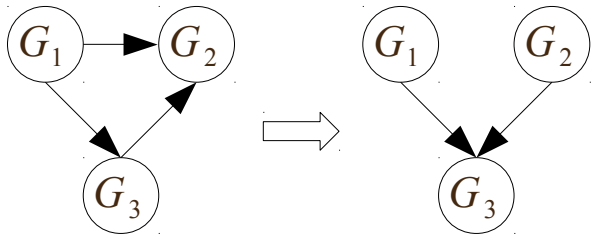
- Comparaison des structures non-orientées via les distances d'édition moyennes sur les 100 jeux de données

- **SGS** et **LAGD** : meilleur score atteint sur 10 répétitions (r=10)

	Alarm		*Insurance*		*Hailfinder*		*Pigs*	
	500	5 000	500	5 000	500	5 000	500	5 000
SGS	~~11~~ **9***	~~8~~ **4***	~~24~~ **24***	~~10~~ **9***	~~41~~ 40	~~29~~ **26***	~~32~~ **1***	~~41~~ 2
LAGD	15	10	**24***	16	47	39	n/a	n/a
GES	11	6	25	15	**39***	33	9	**0***

*** meilleur résultat**

- Correction des v-structures couvertes



Conclusions

- Nous avons proposé deux nouveaux opérateurs locaux pour l'apprentissage de structure de RB.

  - Amélioration des scores atteints à l'aide de l'algorithme SGS
  - Les distances d'édition peuvent être améliorées à l'aide de post-traitements

→ *New Local Move Operators for Learning the Structure of Bayesian Networks.* ***ECAI'12 workshop, AIGM, Montpellier, 2012.***

→ *New Local Move Operators for Bayesian Network Structure Learning.* ***Workshop on Probabilistic Graphical Models, Spain, 2012***.

Plan de l'exposé :

1 – Apprentissage de la structure d'un réseau bayésien

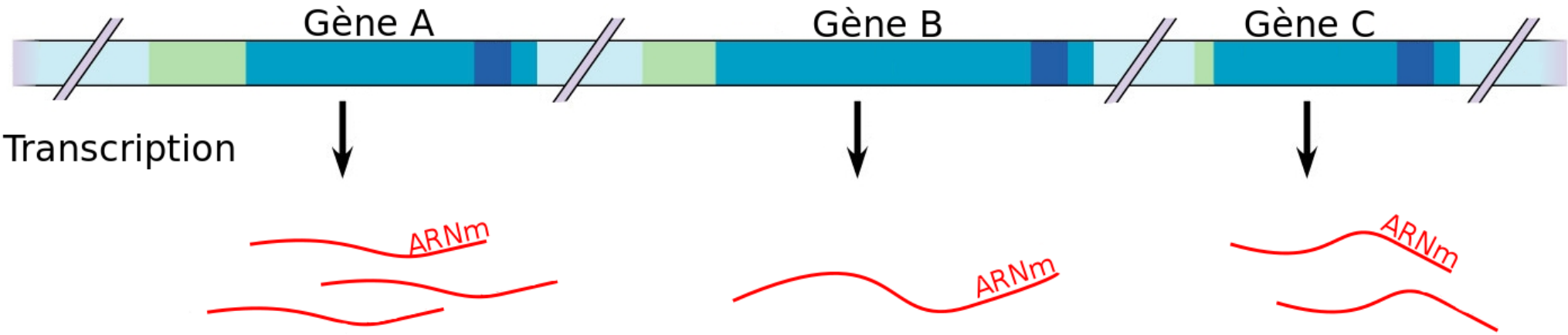
- Présentation des réseaux bayésiens
- État de l'art
- Nos propositions de nouveaux opérateurs locaux

2 – Application à la génétique-génomique

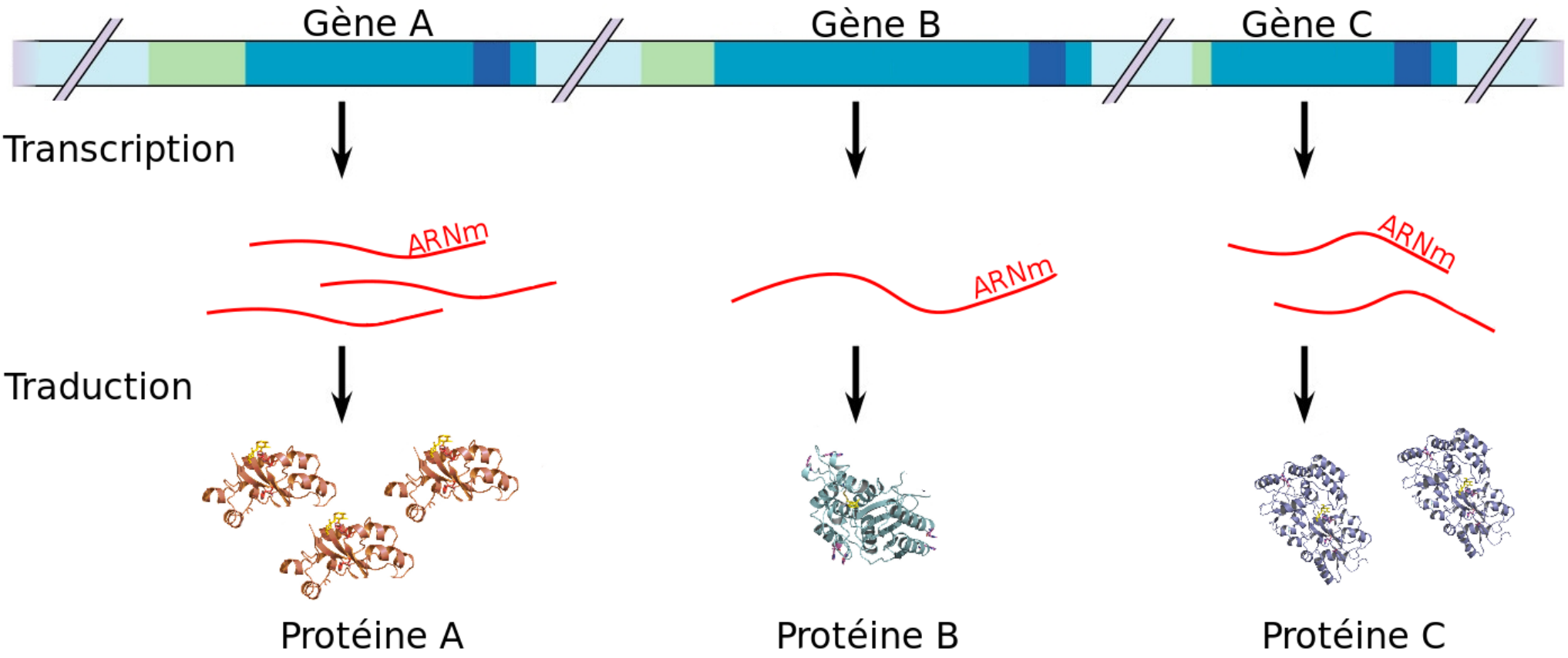
- **Introduction aux réseaux de régulation de gènes**
- **État de l'art**
- **Notre modélisation des données de génétique-génomique**
- **Application aux données d'Arabidopsis *thaliana***

Régulation et polymorphisme

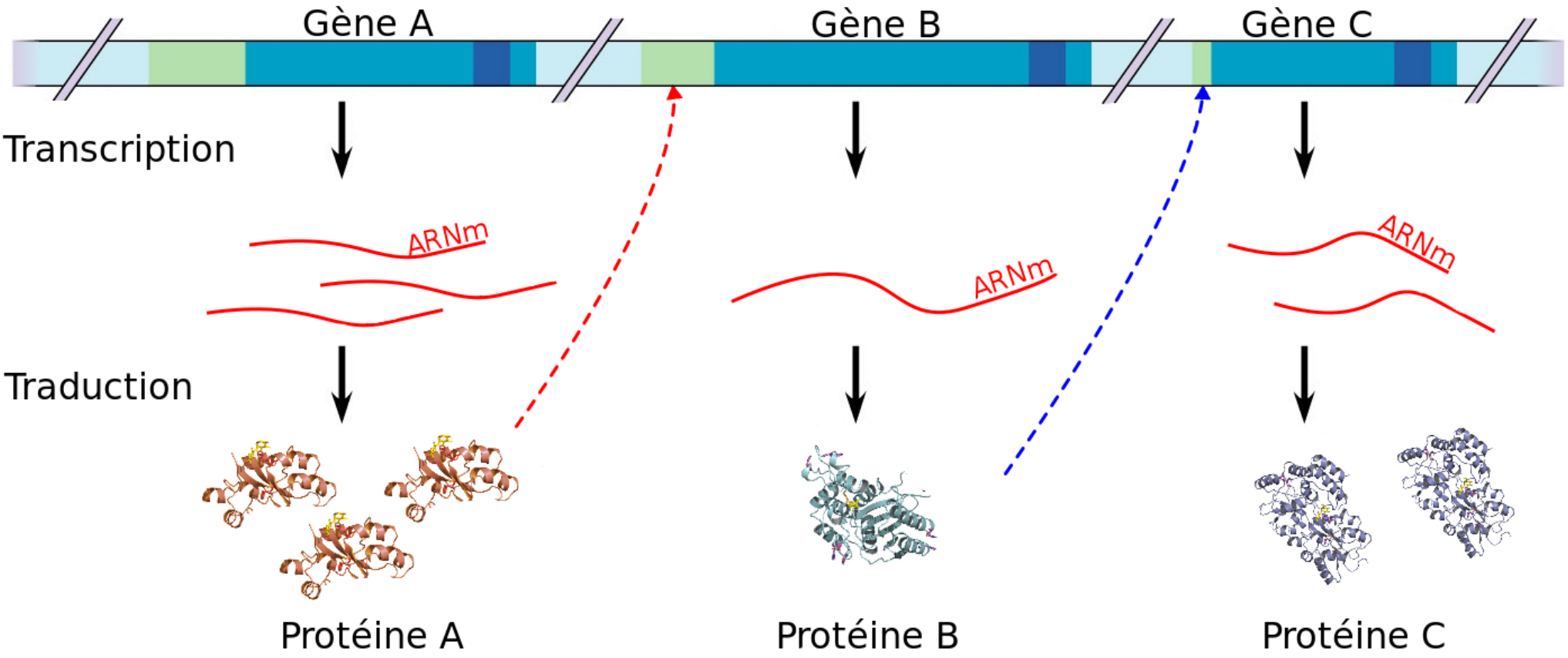
Régulation et polymorphisme



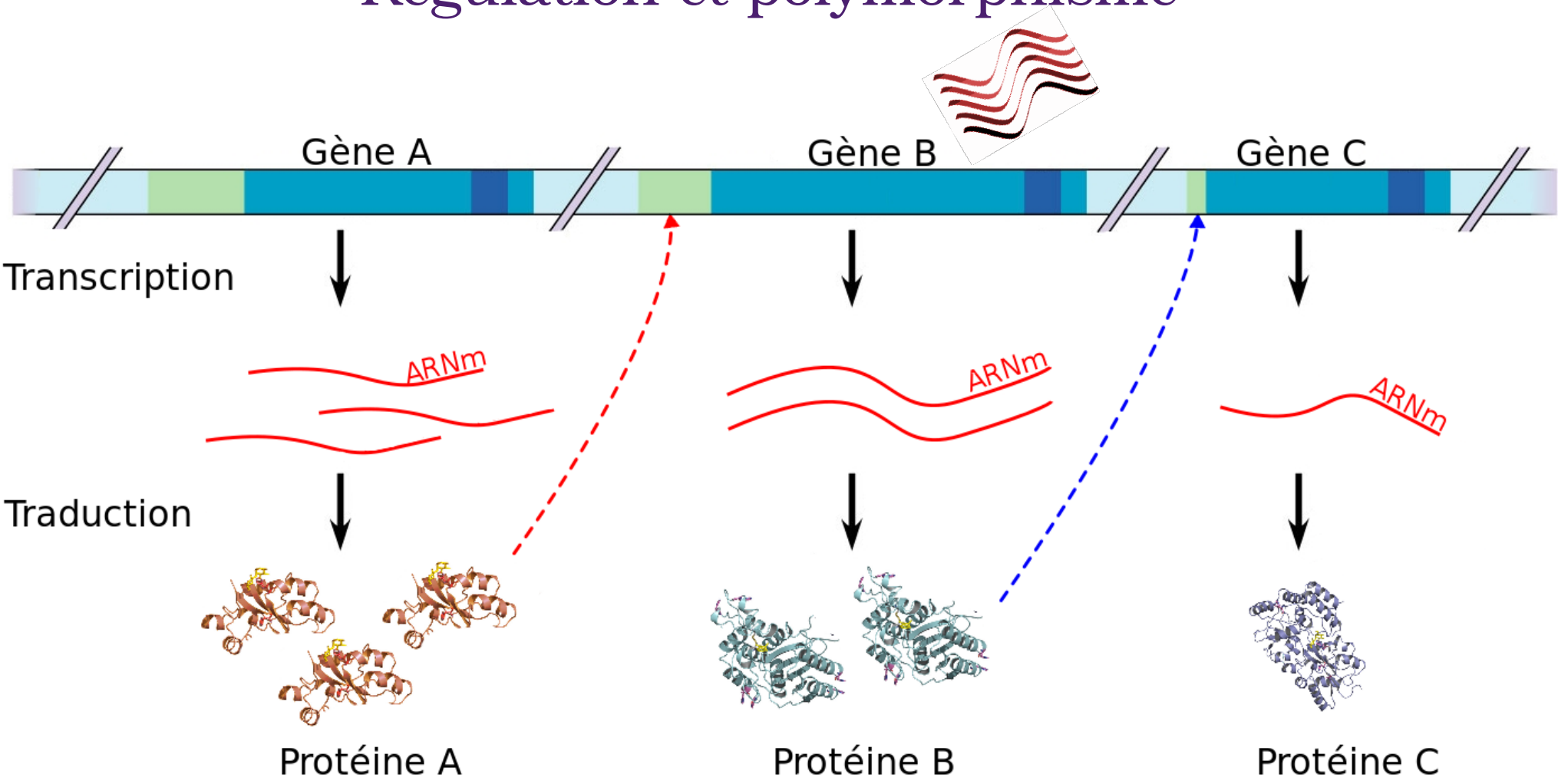
Régulation et polymorphisme



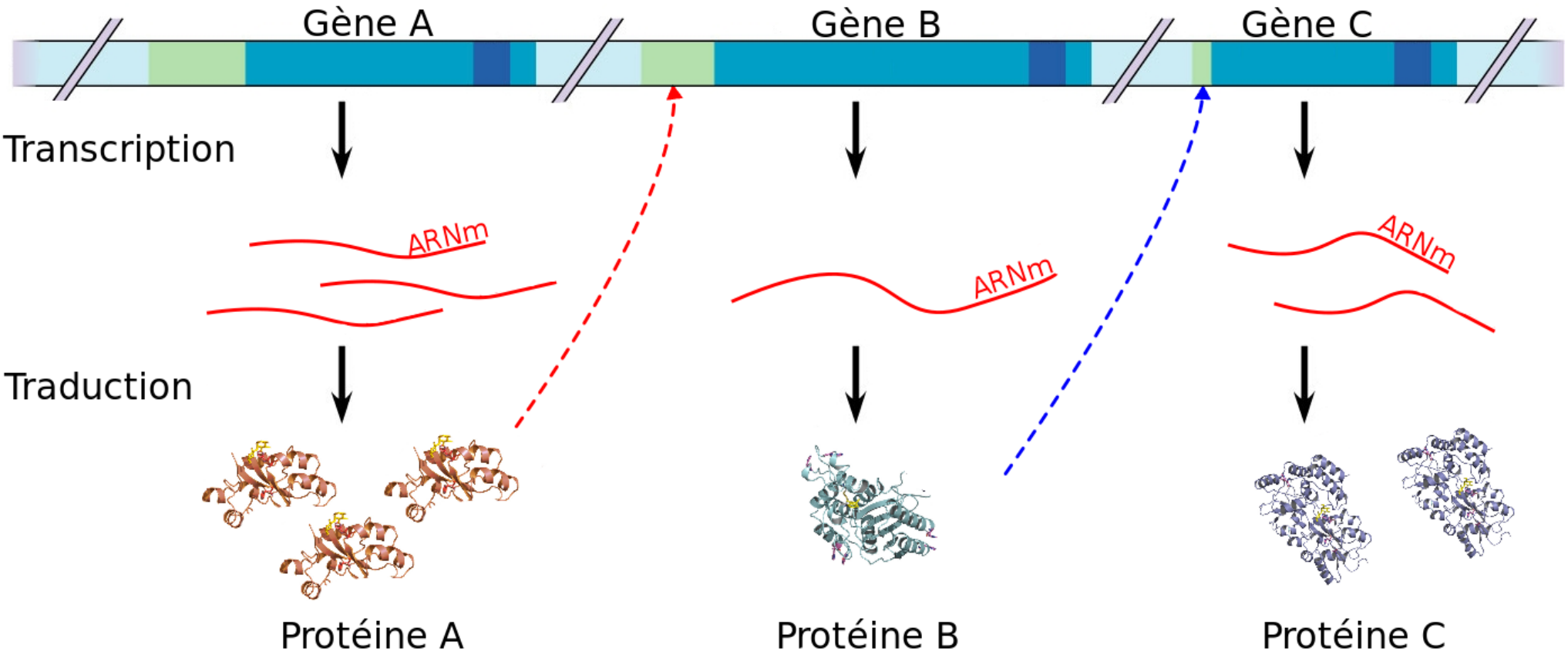
Régulation et polymorphisme



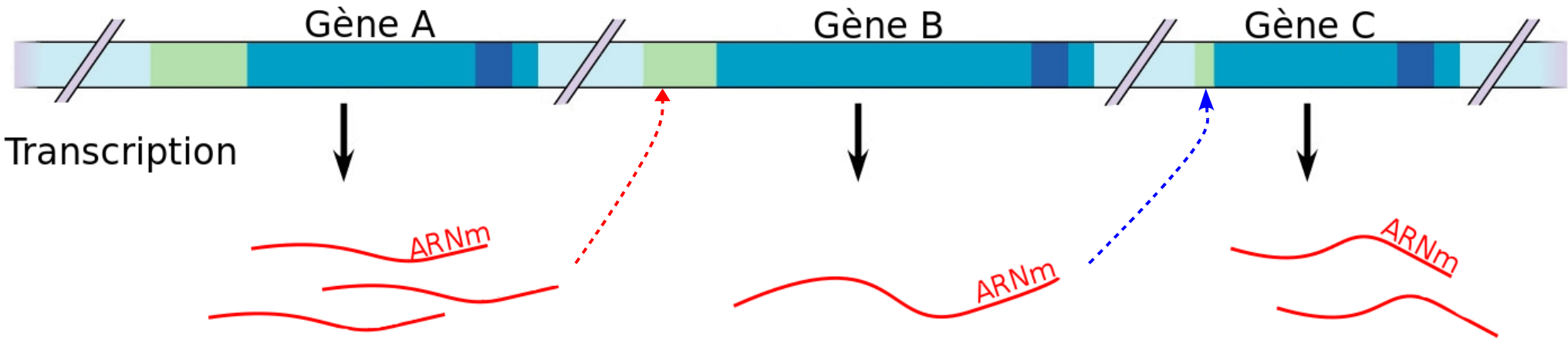
Régulation et polymorphisme



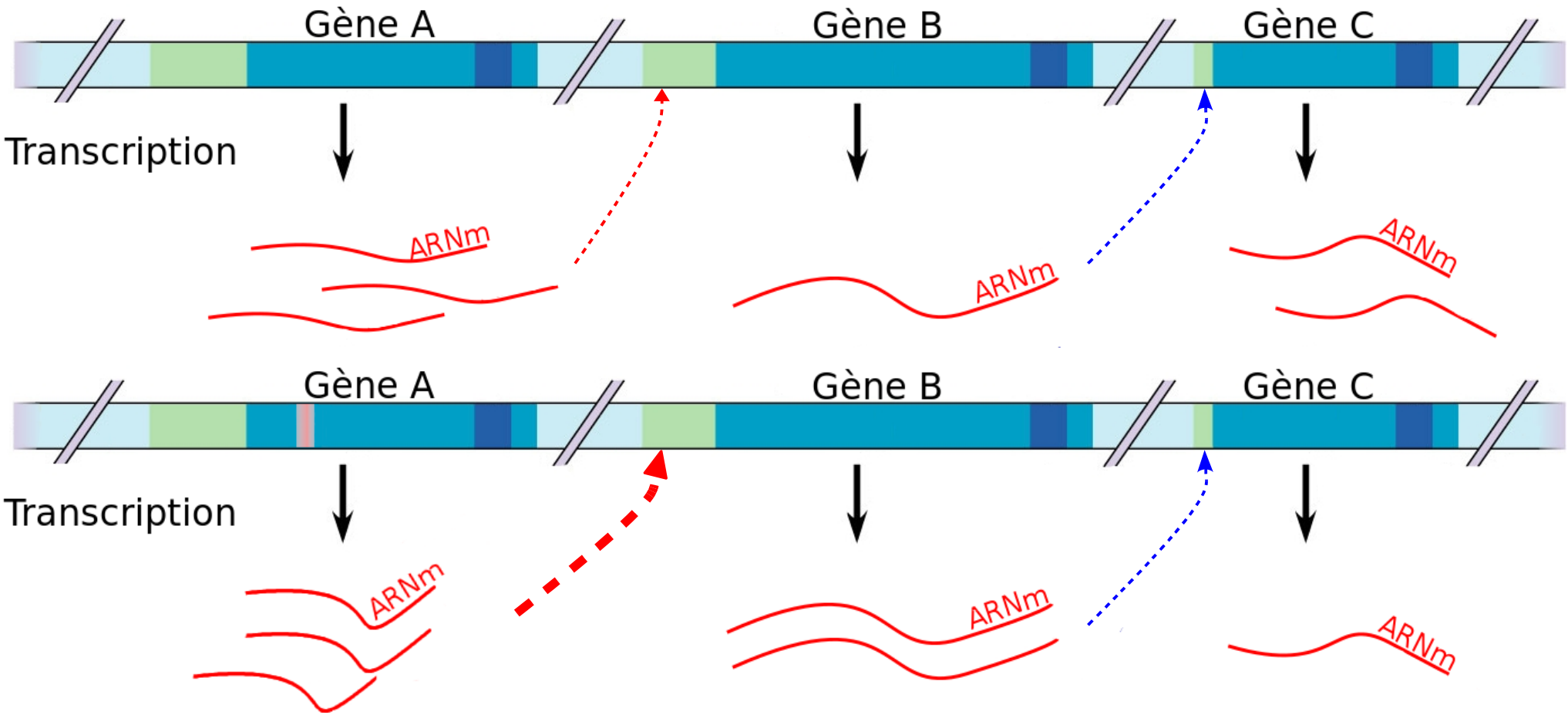
Régulation et polymorphisme



Régulation et polymorphisme

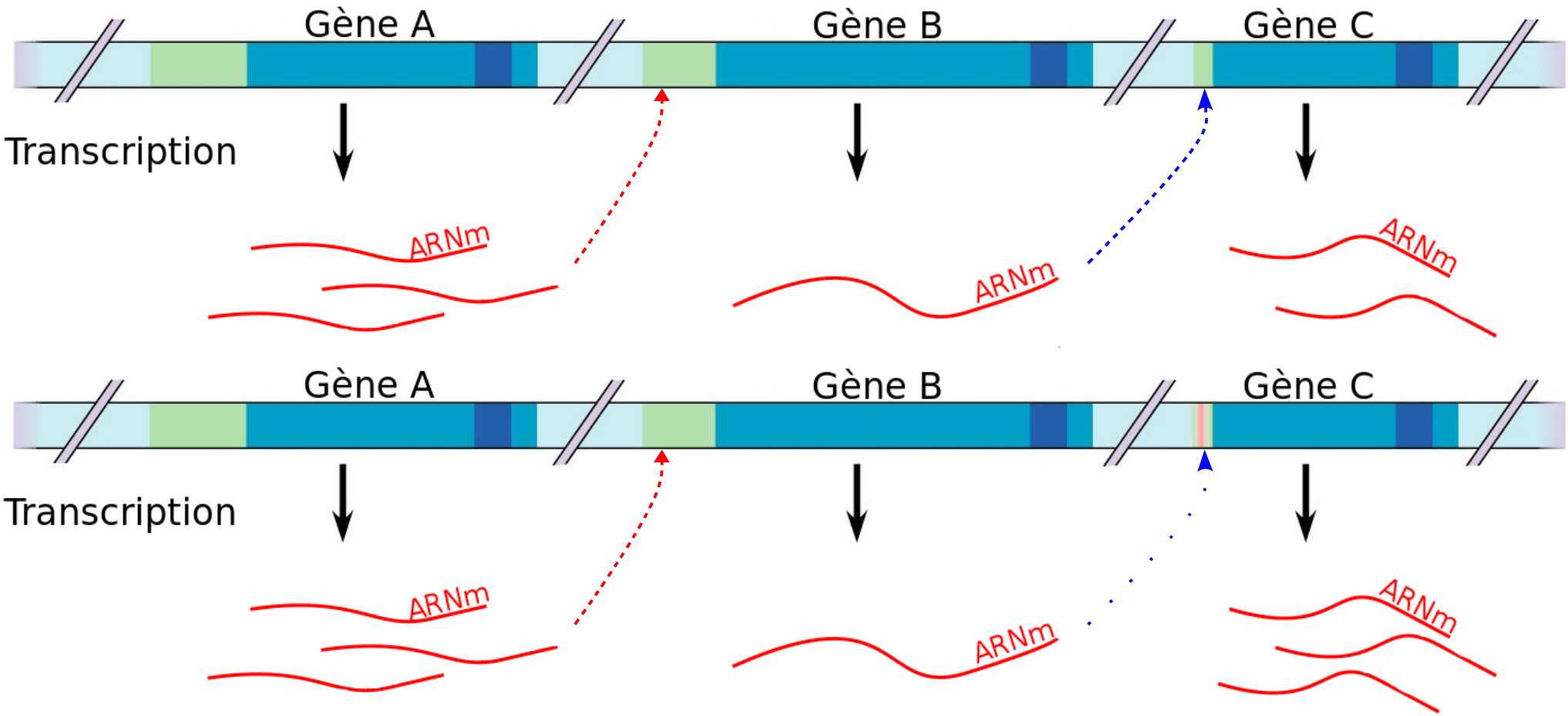


Régulation et polymorphisme



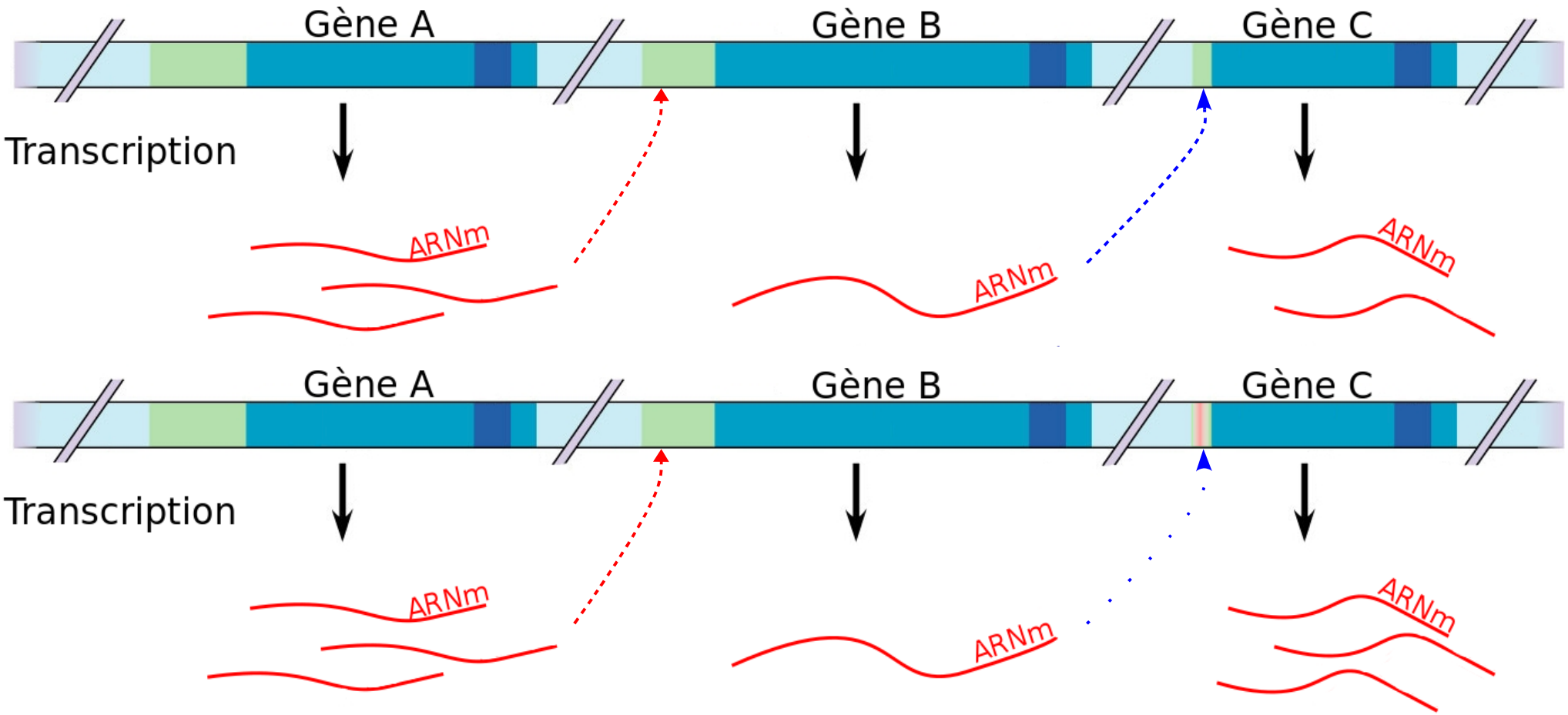
- Mutations de l'ADN : **en région codante → impacte la régulation**

Régulation et polymorphisme



- Mutations de l'ADN : en région codante → impacte la régulation
  en région promotrice → impacte l'expression

Régulation et polymorphisme



- Mutations de l'ADN : en région codante → impacte la régulation
  en région promotrice → impacte l'expression
- Données génétiques : un marqueur par gène (cas idéal)

Apprentissage de RRG – Logiciels disponibles

- Corrélations partielles
  - ARACNE *(Margolin et al., 06)*
  - CLR *(Faith et al., 07)*
  - ParCorA *(de la Fuente et al., 04)*

Données d'expression seules

Données de génétique-génomique

- Forêts aléatoires
  - GENIE3 *(Huynh-Thu et al., 10)*

- Modèles Graphiques Gaussiens (GMM)
  - GGMselect *(Giraud et al., 09)*
  - GeneNet *(Schäfer et al., 05)*
  - Paquet R Lars (régressions Lasso) *(Meinshausen et al., 06)* ***(Vignes et al., 11)***
  - Paquet C glpk (sélecteur de Dantzig) *(Candes et al., 07)* ***(Vignes et al., 11)***
  - SIMoNe *(Chiquet et al., 09)*

- Réseaux bayésiens
  - Banjo *(Hartemink, 05)* ***(Vignes et al., 11) (Vandel et al., à paraître)***
  - SCT *(Chipman et al., 11)*

Modélisation

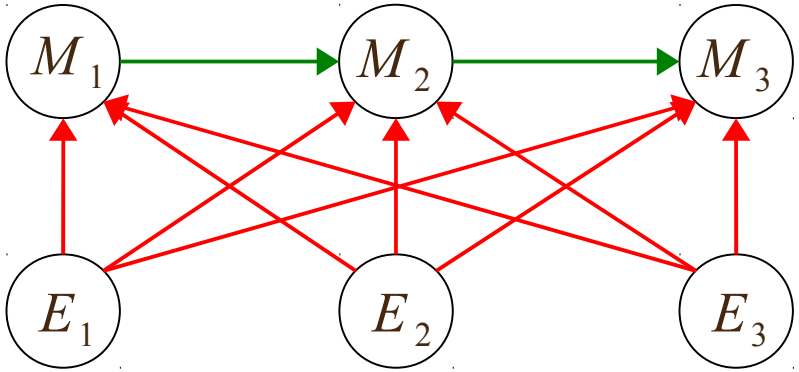
Données d'expression $E_i = \{1,2,3\}$

Modélisation

Données génétiques $M_i = \{0,1\}$

Données d'expression $E_i = \{1,2,3\}$

Modélisation



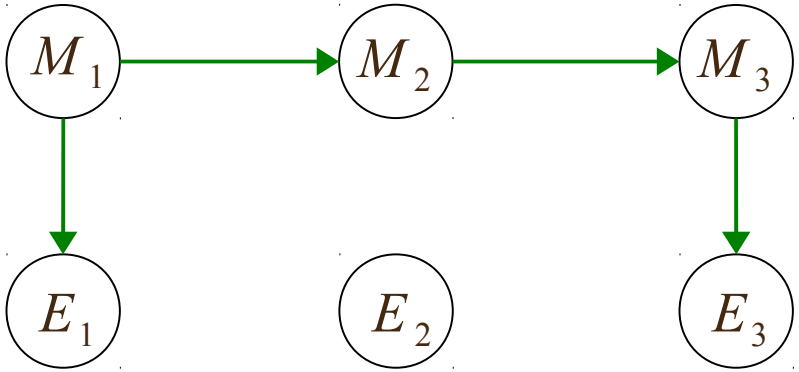
Données génétiques $M_i = \{0,1\}$

Données d'expression $E_i = \{1,2,3\}$

→ Restrictions biologiques

- Impose les arcs Mi → Mi+1
- Interdit les arcs E → M

Modélisation



Données génétiques $M_i = \{0,1\}$

Données d'expression $E_i = \{1,2,3\}$

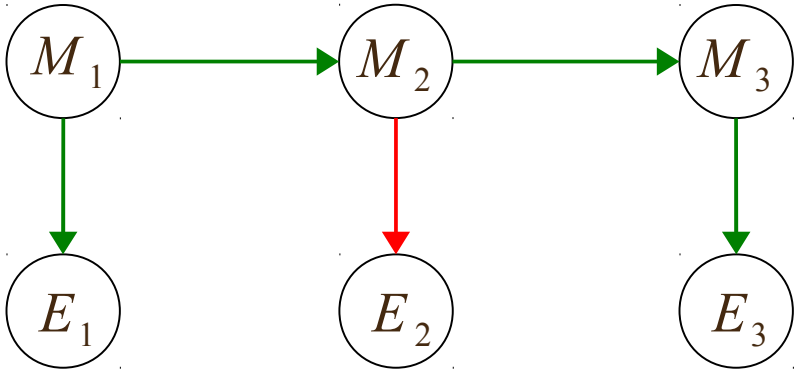
→ Restrictions biologiques

- Impose les arcs $M_i \rightarrow M_{i+1}$
- Interdit les arcs E → M

→ Effet *cis* : mutation en région promotrice du gène i *(exemple : M_1 et M_3)*

- Impose l'arc $M_i \rightarrow E_i$

Modélisation



Données génétiques $M_i = \{0,1\}$

Données d'expression $E_i = \{1,2,3\}$

→ Restrictions biologiques

- Impose les arcs $M_i \rightarrow M_{i+1}$
- Interdit les arcs E → M

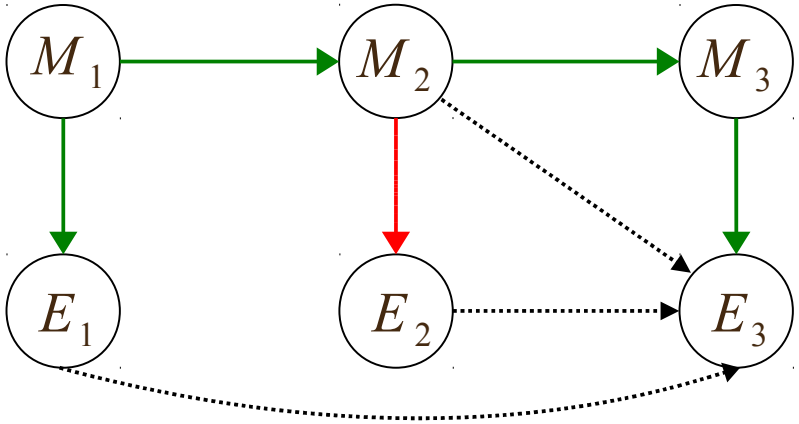
→ Effet *cis* : mutation en région promotrice du gène i *(exemple : M_1 et M_3)*

- Impose l'arc $M_i \rightarrow E_i$

→ Effet *trans* : mutation dans la région codante du gène i *(exemple : M_2)*

- Interdit l'arc $M_i \rightarrow E_i$

Modélisation



Données génétiques $M_i = \{0,1\}$

Données d'expression $E_i = \{1,2,3\}$

→ Restrictions biologiques

- Impose les arcs $M_i \rightarrow M_{i+1}$
- Interdit les arcs E → M

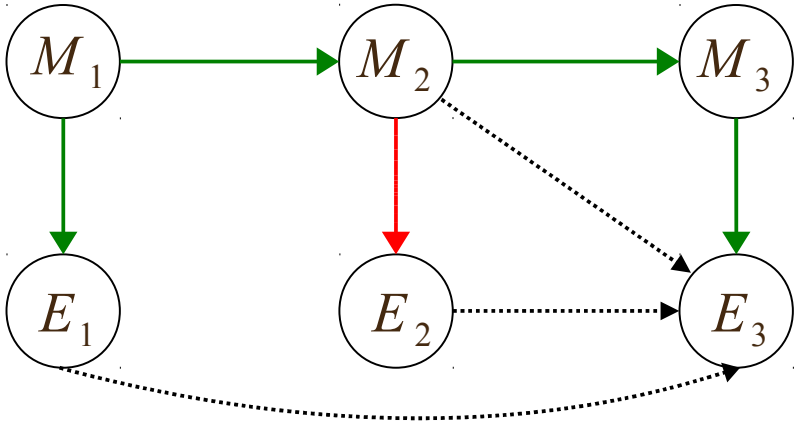
→ Effet *cis* : mutation en région promotrice du gène i *(exemple : M_1 et M_3)*

- Impose l'arc $M_i \rightarrow E_i$

→ Effet *trans* : mutation dans la région codante du gène i *(exemple : M_2)*

- Interdit l'arc $M_i \rightarrow E_i$

Modélisation



Données génétiques $M_i = \{0,1\}$

Données d'expression $E_i = \{1,2,3\}$

→ Restrictions biologiques

- Impose les arcs $M_i \rightarrow M_{i+1}$
- Interdit les arcs E → M

→ Effet *cis* : mutation en région promotrice du gène i *(exemple : M_1 et M_3)*

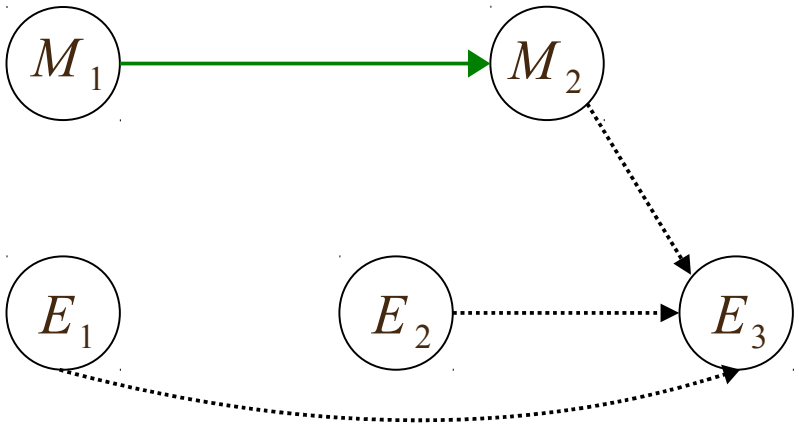
- Impose l'arc $M_i \rightarrow E_i$

→ Effet *trans* : mutation dans la région codante du gène i *(exemple : M_2)*

- Interdit l'arc $M_i \rightarrow E_i$

- Discrétisation adaptative des données d'expression pour une interprétation visuelle

Modélisation



Données génétiques $M_i = \{0,1\}$

Données d'expression $E_i = \{1,2,3\}$

→ Restrictions biologiques

- Impose les arcs $M_i \rightarrow M_{i+1}$
- Interdit les arcs E → M

Situation réelle

nb. de gènes ≠ nb. de marqueurs

→ Effet *cis* : mutation en région promotrice du gène i *(exemple : M_1 et M_3)*

- Impose l'arc $M_i \rightarrow E_i$

→ Effet *trans* : mutation dans la région codante du gène i *(exemple : M_2)*

- Interdit l'arc $M_i \rightarrow E_i$

- Discrétisation adaptative des données d'expression pour une interprétation visuelle

Génération des données

- Données génétiques : populations générées par rétro-croisement avec *CARTHAGENE* *(Schiex et al., 01)*

- Équation Différentielle Ordinaire (EDO) $\frac{dE_i}{dt} = V_i * \prod (Z_j * (\frac{K_{ij}}{I_j + K_{ij}})) * \prod (Z_k (1 + \frac{A_k}{A_k + K_{ak}})) - k_i E_i + \theta E_i$ *(Liu et al., 08)*

Génération des données

- Données génétiques : populations générées par rétro-croisement avec *CARTHAGENE* *(Schiex et al., 01)*

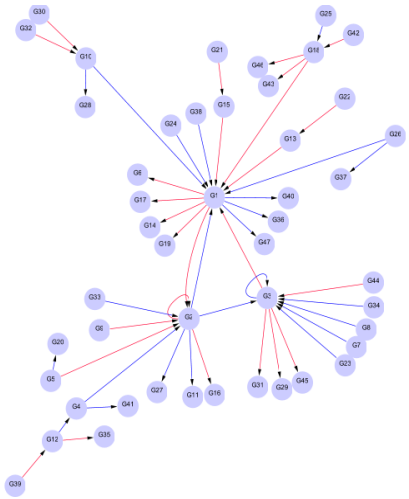
- Équation Différentielle Ordinaire (EDO)

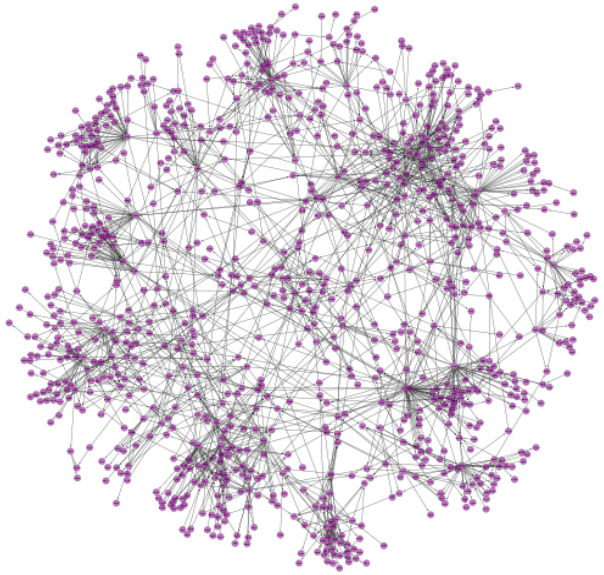
$$\frac{dE_i}{dt} = V_i * \prod (Z_j * (\frac{K_{ij}}{I_j + K_{ij}})) * \prod (Z_k (1 + \frac{A_k}{A_k + K_{ak}})) - k_i E_i + \theta E_i$$

(Liu et al., 08)

50 gènes
50 → 500 individus

1000 gènes
100 → 999 individus





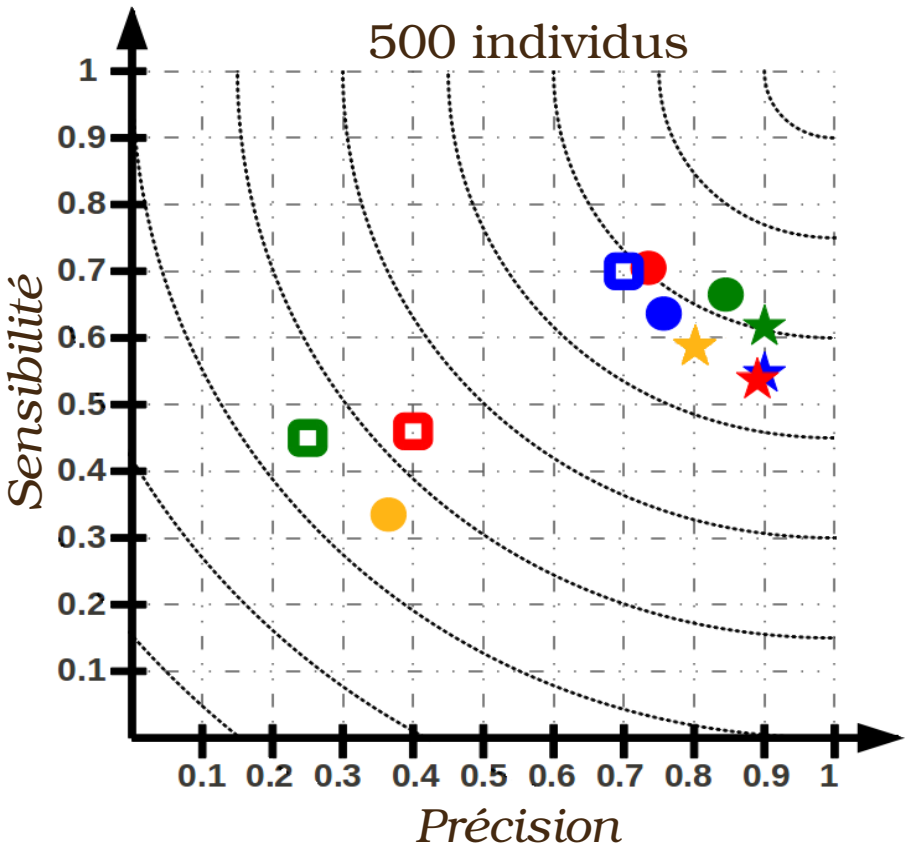
50 réseaux Web50 *(Mendes et al., 03)*

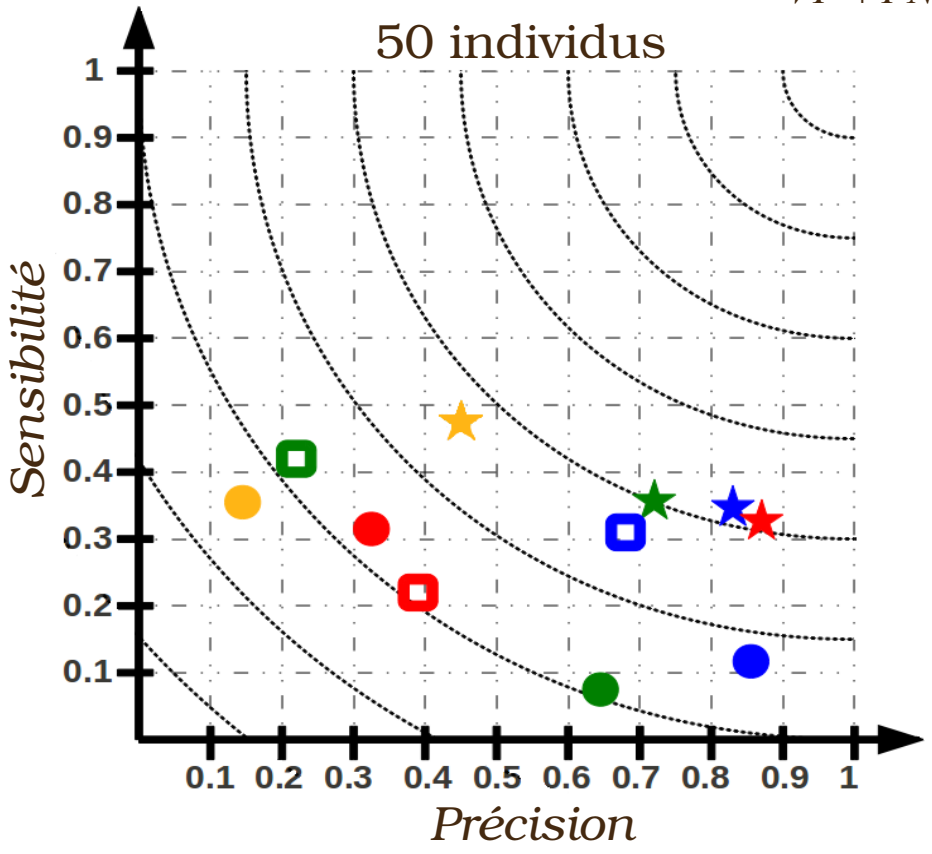
15 réseaux DREAM *(Pinna et al., 11)*

Réseaux Web50

$$Précision : \frac{VP}{VP+FP}$$

$$Sensibilité : \frac{VP}{VP+FN}$$





Moyennes sur les 50 réseaux Web50 non-orientés

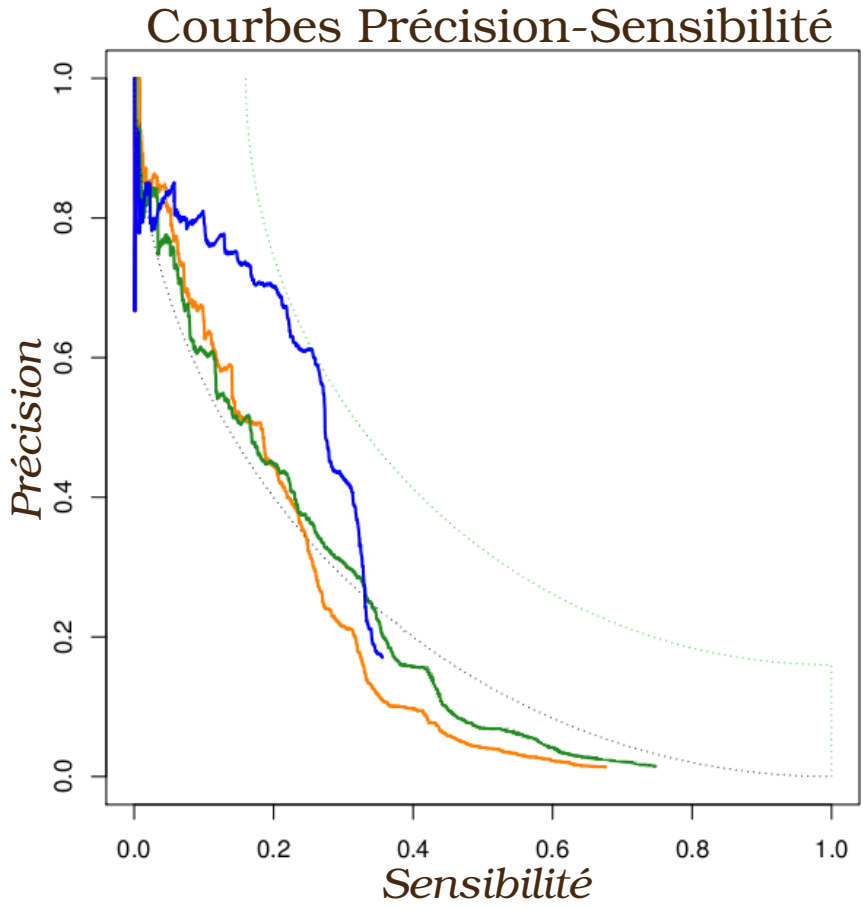
- **Réseaux bayésiens** : GS (BDeu ★ / BIC ★ / fNML ★), SCT ★
- **Modèle linéaire** : Régressions Lasso ●, GGMselect ●, GeneNet ●, SIMoNe ●
- **Corrélation de paires** : CLR □, ARACNE □, ParCorA □

Réseaux DREAM

Fréquences d'apparition pondérées des arcs sur 20 exécutions

G_{725}	G_{321}	1
G_{910}	G_{18}	0.98
G_{221}	G_{119}	0.95
G_{901}	G_{544}	0.92
G_{748}	G_{92}	0.89
G_{221}	G_{226}	0.88
⋮		

Réseaux orientés



- Réseaux bayésiens (GS score BDeu)
- Régressions Lasso
- Sélecteur de Dantzig

Réseau 6
300 individus
Connectivité~2

Arabidopsis thaliana

- **Données expérimentales** *(Simon et al., 08)*
  - Population RIL 158 individus (Cvi/Col)
  - Puces CATMA 34660 sondes (22089 gènes)
  - Marqueurs SNP 89

Arabidopsis thaliana

- **Données expérimentales** *(Simon et al., 08)*
  - Population RIL 158 individus (Cvi/Col)
  - Puces CATMA 34660 sondes (22089 gènes)
  - Marqueurs SNP 89

- **Pré-traitement des données d'expressions**
  - complétion des manquants à l'aide de sondes prédictrices
  - sélection de 4176 sondes expliquées par le polymorphisme (eQTL)

Arabidopsis thaliana

- **Données expérimentales** *(Simon et al., 08)*
  - Population RIL 158 individus (Cvi/Col)
  - Puces CATMA 34660 sondes (22089 gènes)
  - Marqueurs SNP 89

- **Pré-traitement des données d'expressions**
  - complétion des manquants à l'aide de sondes prédictrices
  - sélection de 4176 sondes expliquées par le polymorphisme (eQTL)

- **Pré-traitement des génotypes**
  - complétion des manquants avec le paquet R « qtl »
  - création de pseudos-marqueurs (590 marqueurs au total)

Arabidopsis thaliana

- **Données expérimentales** *(Simon et al., 08)*
  - Population RIL 158 individus (Cvi/Col)
  - Puces CATMA 34660 sondes (22089 gènes)
  - Marqueurs SNP 89

- **Pré-traitement des données d'expressions**
  - complétion des manquants à l'aide de sondes prédictrices
  - sélection de **4176 sondes** expliquées par le polymorphisme (eQTL)

- **Pré-traitement des génotypes**
  - complétion des manquants avec le paquet R « qtl »
  - création de pseudos-marqueurs (**590 marqueurs** au total)

nb. de gènes ≠ nb. de marqueurs

Arabidopsis thaliana

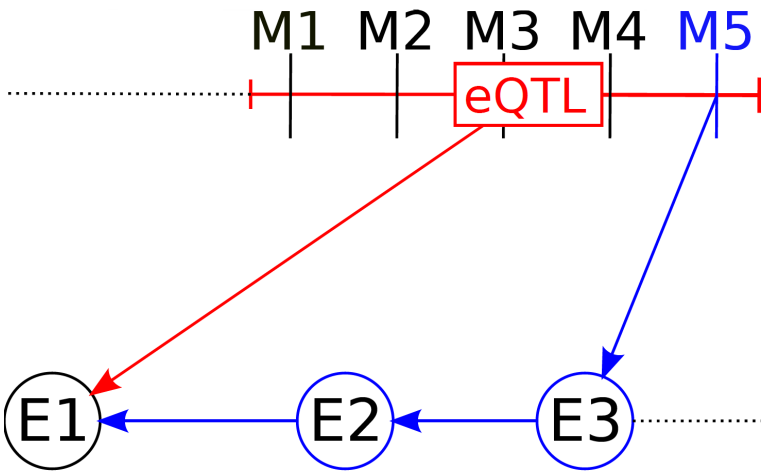
- Réseau bayésien appris
  - 4766 variables
  - 6137 arcs (284 $M_i \rightarrow E_j$ / 5853 $E_i \rightarrow E_j$)

Arabidopsis thaliana

- Réseau bayésien appris
  - 4766 variables
  - 6137 arcs (284 $M_i \rightarrow E_j$ / 5853 $E_i \rightarrow E_j$)

- Comparaison avec *une analyse eQTL **
  ** recherche pour chaque gène, les polymorphismes expliquant son expression*

  → ~48% des eQTL detectés en *trans* sont expliqués par un chemin dans notre RB

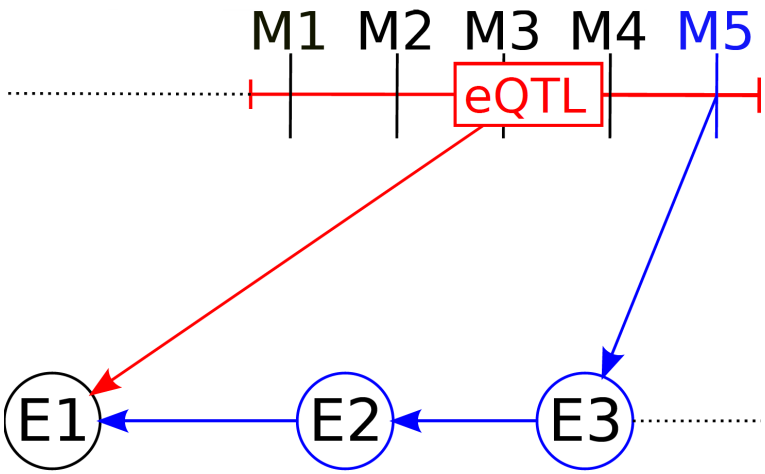


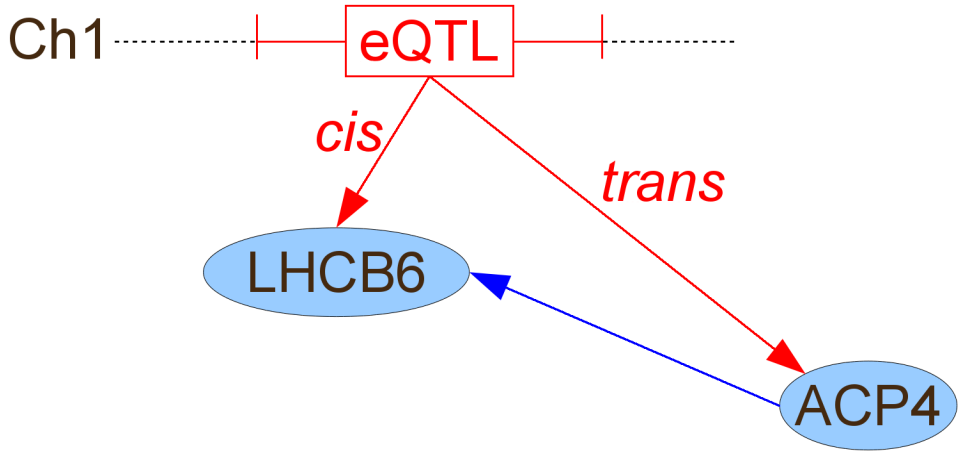
Arabidopsis thaliana

- Réseau bayésien appris
  - 4766 variables
  - 6137 arcs (284 $M_i \rightarrow E_j$ / 5853 $E_i \rightarrow E_j$)

- Comparaison avec *une analyse eQTL **
  ** recherche pour chaque gène, les polymorphismes expliquant son expression*

  → ~48% des eQTL detectés en *trans* sont expliqués par un chemin dans notre RB





→ *LHCB6 & ACP4 impliqués dans le phénomène de photosynthèse chez A. thaliana*

Conclusions

- Nous avons proposé une modélisation par RB dans le cadre des données de génétique-génomique.
  - Approche compétitive sur des réseaux de petite taille (Web50) et pour des réseaux plus larges (DREAM)

- Nous avons reconstruit un premier RRG pour *Arabidopsis thaliana* à partir de données réelles.
  - Résultats cohérents avec une analyse eQTL
  - Validation avec la littérature de certaines régulations apprises

→ *Inférence de réseaux de régulation de gènes au travers de scores étendus dans les réseaux bayésiens.* **Revue d'Intelligence Artificielle, à paraître.**

→ *Gene Regulatory Network Reconstruction Using Bayesian Networks, the Dantzig Selector, the Lasso and Their Meta-Analysis.* **PLoS ONE, 2011.**

Conclusions générales

- Les réseaux bayésiens représentent une approche performante pour l'apprentissage de RRG
  - Modélisation d'effets non-linéaires
  - Évaluation globale du réseau
  - Représentation de la causalité induite par les données de génétique-génomique
- Amélioration de l'apprentissage des RB à l'aide de voisinages étendus
  - Proposition d'opérateurs locaux : SWAP et extension itérative
  - Post-traitement de la structure apprise
- Questions en suspens
  - Approximations dues : - à l'absence de circuit
    - à la discrétisation
    - au caractère glouton de notre approche

Perspectives à court terme

- Apprentissage de réseaux bayésiens
  - Enrichir notre comparatif d'autres approches : *Optimal Reinsertion*
  - Tester les opérateurs avec des heuristiques plus évoluées : *Tabu*
  - Comparer les réseaux appris au réseau optimal
  - Diffuser l'algorithme SGS sur des logiciels existants (*Cytoscape*)

- Application aux données de génétique-génomique
  - Évaluer les méthodes de discrétisation
  - Poursuivre l'étude du RRG d'*Arabidopsis thaliana*

Perspectives à plus long terme

- Apprentissage de réseaux bayésiens
  - Continuer d'améliorer la distance d'édition à l'aide de post-traitements

- Application aux données de génétique-génomique
  - Enrichir le modèle de nouveaux types de données
  - Restreindre intelligemment l'espace de recherche

Références

Publications :

« *Inférence de réseaux de régulation de gènes au travers de scores étendus dans les réseaux bayésiens* », Revue d'Intelligence Artificielle (RIA), A paraitre.

« *Gene Regulatory Network Reconstruction Using Bayesian Networks, the Dantzig Selector, the Lasso and Their Meta-Analysis* », PLoS ONE, 2011.
http://www.inra.fr/mia/T/degivry/Vignes11a.pdf

Présentations orales :

« *New Local Move Operators for Bayesian Network Structure Learning* »
Workshop on Probabilistic Graphical Models, Spain, 2012.
http://www.inra.fr/mia/T/degivry/Vandel12c.pdf

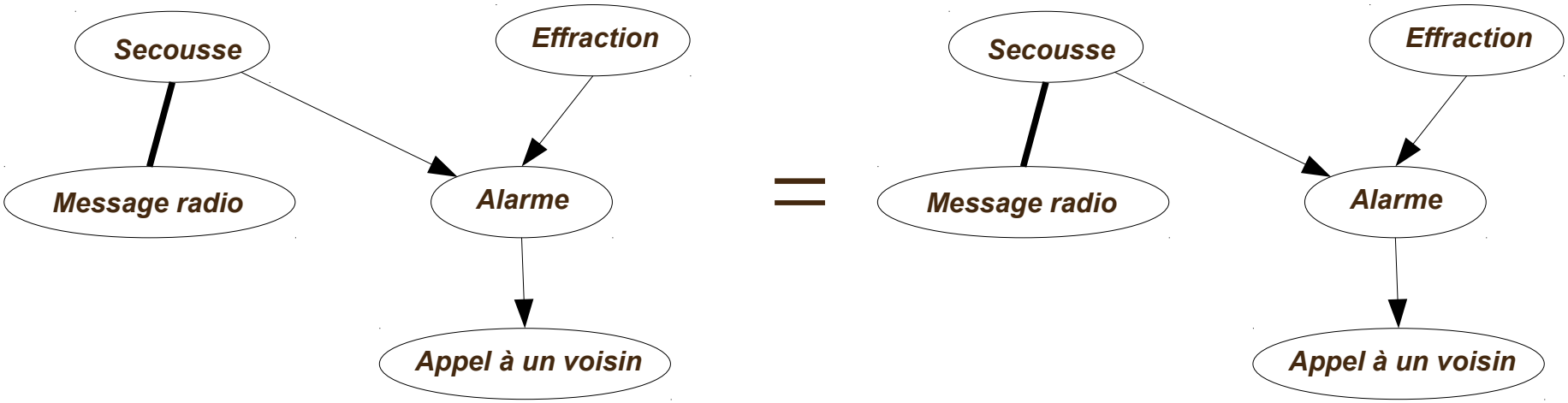
« *New Local Move Operators for Learning the Structure of Bayesian Networks* »
ECAI'12 workhop, Algorithmic issues for inference in graphical models, Montpellier, 2012.
http://www.inra.fr/mia/T/degivry/Vandel12b.pdf

« *A New Local Move Operator for Reconstructing Gene Regulatory networks* »
CP'11 workshop, Constraint Based Methods for Bioinformatics, Italy, 2011.
http://www.inra.fr/mia/T/degivry/Vandel11a.pdf

« *Extended bayesian scores for reconstructing gene regulatory networks* »
ECCS'10 workshop, Graphical models for reasoning on biological systems, Portugal, 2010.
http://www.inra.fr/mia/T/degivry/Vandel10b.pdf

« *Reconstruction de réseau de régulation de gène à l'aide de données génomiques et de données génétiques* »
CAp'10, Conférence sur l'apprentissage automatique, Clermont-Ferrand, 2010.
http://www.inra.fr/mia/T/degivry/Vandel10a.pdf

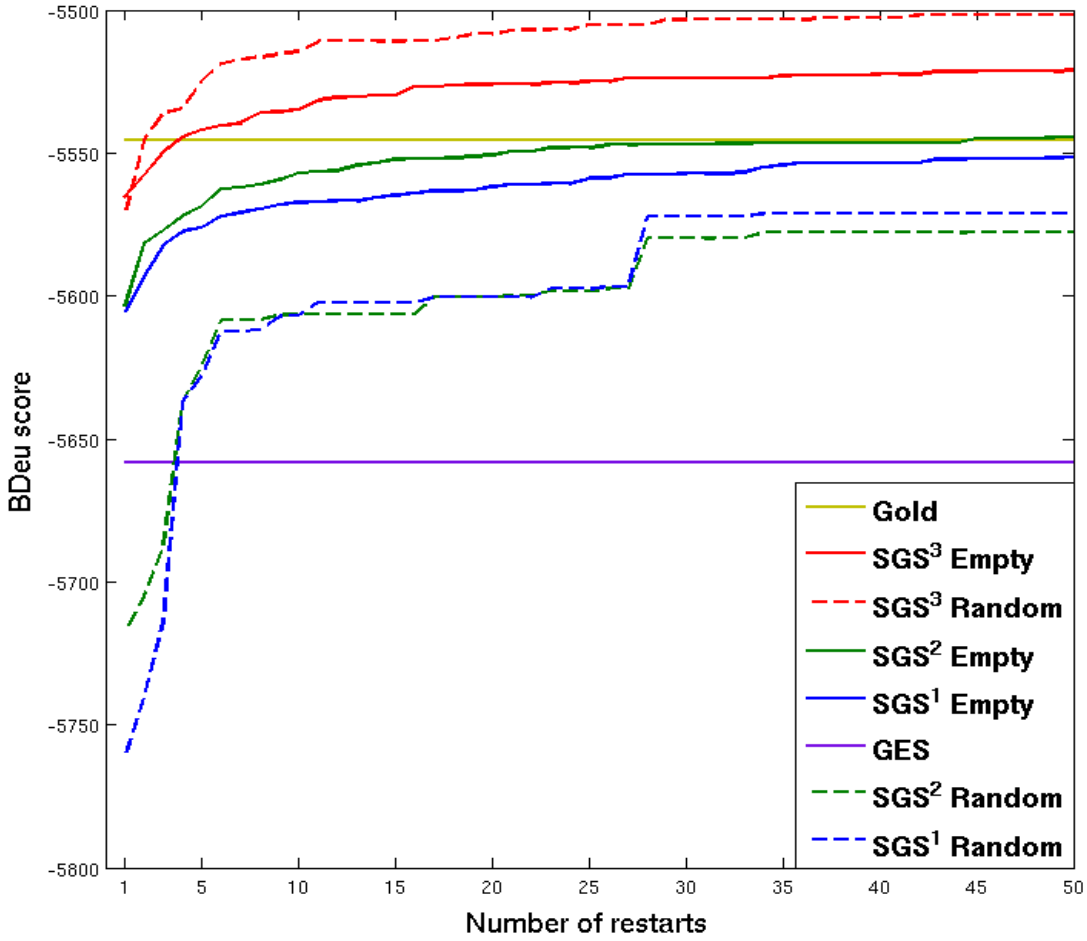
Équivalence de Markov



- Deux instances d'une même classe d'équivalence représentée par un cpDAG
  (**c**ompleted **p**artially **D**irected **A**cyclic **G**raph)

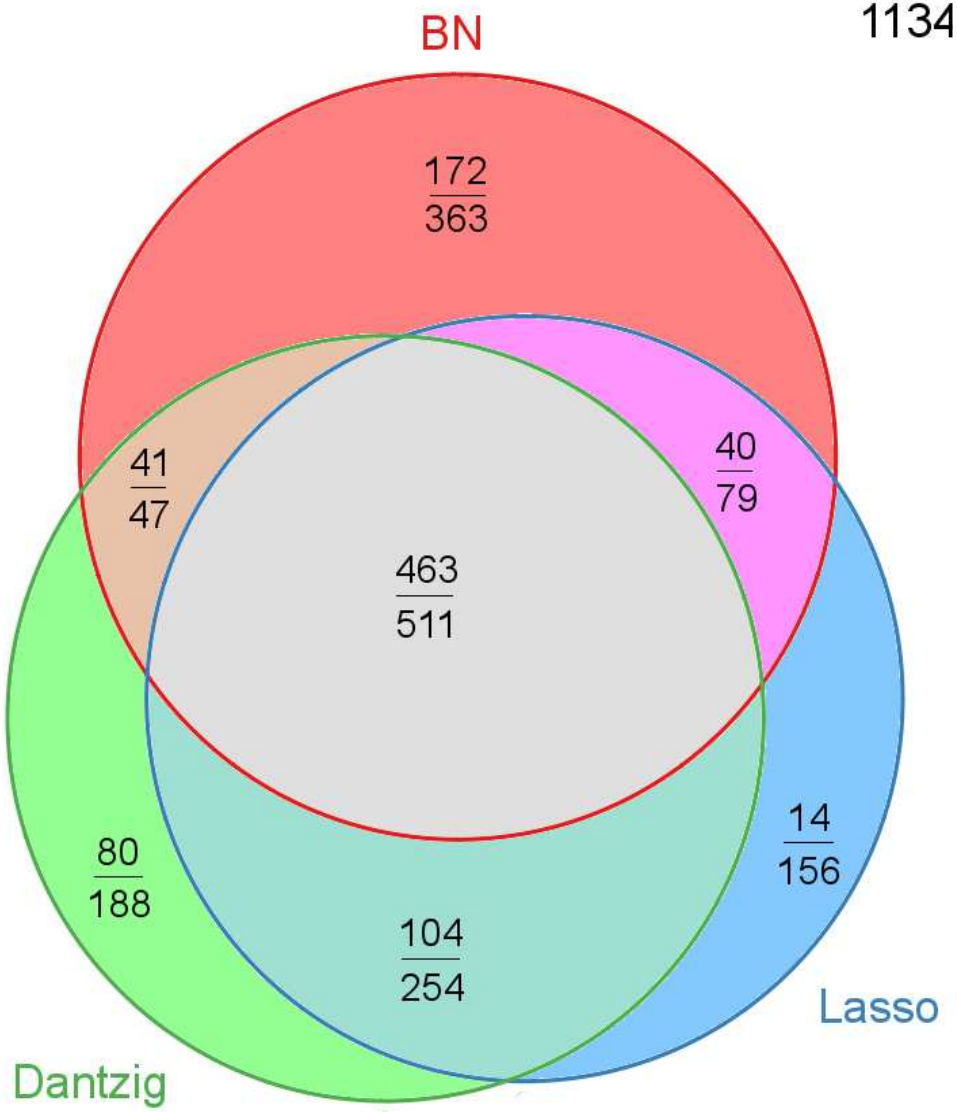
Comparatifs : Nombre de répétitions

- Impact du nombre d'exécutions **r**
- Moyennes sur 30 jeux de données du réseau Alarm (500 exemples)
- SGS initialisé avec le graphe vide et le graphe aléatoire (2 parents max)



Réseaux DREAM

- Nombre d'arcs en commun parmi les 1000 premiers classés par les 3 approches



$$\frac{VP}{VP+FP}$$

Réseau 11
999 individus
connectivité~2