Méthodes d'inférence de réseaux de gènes

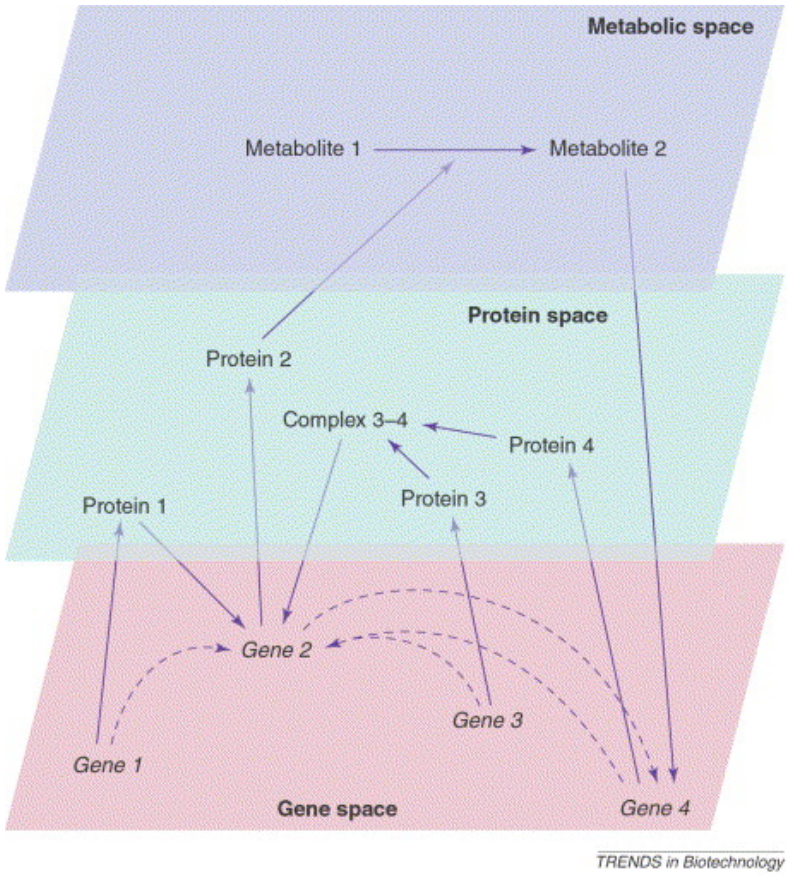
B. Duval - O. Goudet

UA

Janvier 2026

On revient aux réseaux de gènes

Objectif : comprendre les interactions entre les gènes grâce aux méthodes de mesures de l'expression de gènes (puces à ADN, RNA-seq); les interactions seront représentées par des réseaux



Objectifs

- Présenter certaines méthodes utilisées dans la littérature pour l'inférence de réseaux biologiques.
- Présenter les générateurs de données qui permettent d'évaluer les méthodes.

Section 1

Contexte

Inférence de réseaux de gènes

- Données transcriptomiques
  - **Données statiques** : mesures de l'expression des gènes dans un état stable (steady-state)
    Ex: une plante soumise à un stress hydrique
  - **Données temporelles** : mesures de l'expression des gènes sur un intervalle de temps pour suivre l'évolution d'un processus biologique
    Exemple : une graine observée pendant sa période de germination

Dans ce cours, nous ne considérons que des données statiques.
Les données soumises à la méthode d'inférence sont donc sous la forme d'une matrice d'expression de gènes.

Exemple: Inférence d'un graphe de régulation génétique à partir de données d'observation continues

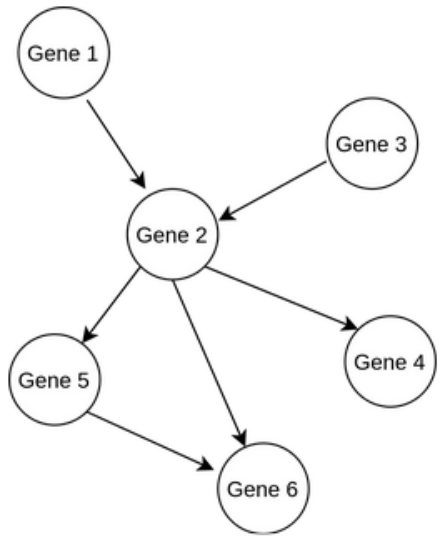
Matrice d'expression du génome de la plante *Arabidopsis thaliana*

26 374 gènes

1042 Expériences

	Gène 1	Gène 2	Gène 3	Gène 4	Gène 5	Gène 6	...
Expérience 1	1,13	1,49	0,78	1,71	2,11	5,09	
Expérience 2	1,10	0,58	1,25	0,37	1,04	4,56	
Expérience 3	2,13	1,40	0,40	1,56	2,03	5,48	
Expérience 4	0,70	0,78	0,85	0,62	1,51	4,25	
Expérience 5	1,52	1,28	0,88	0,85	1,35	5,55	
Expérience 6	2,01	2,04	-0,04	0,10	3,14	4,27	
Expérience 7	0,74	0,54	1,45	0,20	2,52	5,92	
Expérience 8	0,23	0,85	1,61	0,15	2,14	6,04	
Expérience 9	1,64	1,64	1,65	1,52	3,53	6,25	
Expérience 10	1,10	1,37	1,28	0,97	3,02	5,71	
Expérience 11	0,77	0,92	0,83	0,73	3,14	6,16	
Expérience 12	1,39	1,73	0,93	1,24	3,49	5,88	
Expérience 13	1,41	1,65	1,32	1,18	3,88	7,02	
Expérience 14	1,57	2,25	0,61	0,80	1,74	4,80	
Expérience 15	0,98	1,67	0,42	0,13	1,96	4,64	
Expérience 16	2,23	1,71	0,83	0,85	1,91	7,39	
...							

Exemple de graphe de régulation génétique inféré



Une typologie des méthodes utilisées en bioinformatique

- Méthodes basées sur une mesure de co-expression :
  - Mesure de corrélation linéaire
  - Mesure d'information mutuelle
- Méthodes basées sur la régression.

Section 2

Exemple de méthode utilisant la corrélation

WGCNA - Weighted Gene Co-expression Network Analysis

Langfelder and Horvath (2008)

- Méthode qui s'appuie sur la corrélation linéaire pour définir les liens candidats
- Exploite des propriétés topologiques pour affiner le réseau
- S'intéresse aux **modules** du réseau plus qu'aux liens entre paires de gènes.
  Un module est un cluster de gènes fortement interconnectés: lien avec les fonctions biologiques des gènes.

WGCNA - Construction d'un réseau

- Calcul de la matrice de corrélation non signée Cor: valeur absolue de la corrélation de Pearson
- Seuillage de la matrice de corrélation pour obtenir la matrice d'adjacence $\mathcal{A}$ du réseau
  - hard threshold
    pour un seuil τ fixé, $a_{i,j} = 1 \iff Cor_{i,j} > \tau$
  - soft threshold
    pour un paramètre β à déterminer, $a_{gs} = |cor(g, s)|^{\beta}$
- Choix par l'utilisateur d'un seuil approprié pour un réglage approprié précision/rappel.

WGCNA - Recherche de modules

La recherche de modules s'appuie sur la mesure TOM Topological Overlap Matrix

$$\omega_{gs} = \frac{l_{gs} + a_{gs}}{\min(k_g, k_s) + 1 - a_{gs}}$$

où a_{gs} est la mesure d'adjacence entre deux gènes g et s
$l_{gs} = \sum_v a_{gv} a_{vs}$ est la connectivité des voisins communs entre g et s
et $k_g = \sum_v a_{gv}$ est la connectivité du gène g.

La TOM a pour but de quantifier la co-expressivité entre 2 gènes en prenant en compte leur corrélation, mais aussi la corrélation entre leurs voisins

WGCNA - Recherche de modules

Méthode de classification ascendante hiérarchique basée sur la TOM pour détecter les modules

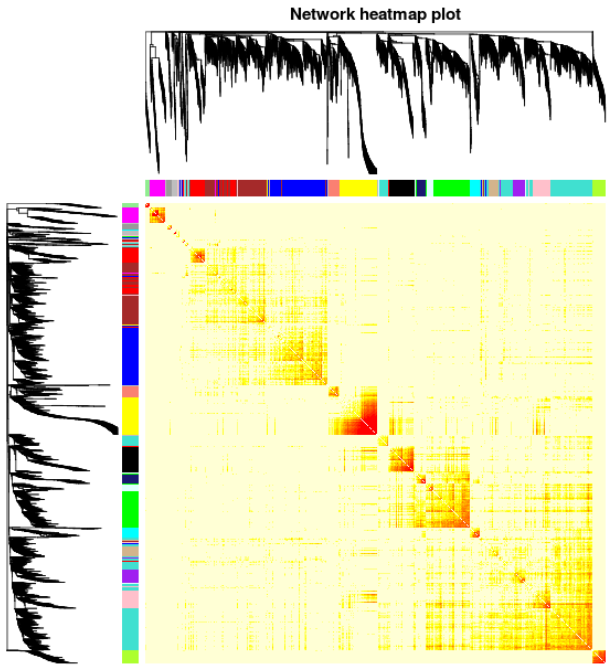


Figure 1: Le réseau est représenté grâce à une heat map de la TOM, avec sur les axes, le dendrogramme dessinant les modules du bloc. L'échelle de coloration va de jaune pour une valeur faible à rouge pour une valeur forte.

WGCNA Complexité et disponibilité

- Complexité
  Pour un jeu de données constitué de n gènes et p échantillons
  - l'estimation de la corrélation est en $\mathcal{O}(n^2 p)$
  - le calcul de la TOM est en $\mathcal{O}(n^3)$
- Code: Disponible sous forme de package R
  https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/

Méthode WGCNA utilisée dans de nombreux travaux en biologie et médecine

- Saris et al. (2009) : analyse de réseau du sang périphérique de patients atteints de sclérose latérale amyotrophique. `https://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-10-405`
- Farhadian et al. (2021) : étude de processus de lactation. `https://www.nature.com/articles/s41598-021-81888-z.pdf`

Extension possible

- Remplacer le calcul de corrélation par un score d'information mutuelle pour étendre la méthode à d'autres données qui rentrent mal dans le cas "linéaire Gaussien" (cf. Cours 2).
- Une autre idée pour le calcul de la mesure TOM ?

Section 3

Exemple de méthode utilisant l'information mutuelle

ARACNE / Algorithm for Reconstruction of Accurate Cellular Network

ARCANE Margolin et al. (2006)

- Matrice de liens basée sur l'Information Mutuelle (IM)
- Elimination de liens indirects

Une méthode en 3 étapes:

1. Calcul de la matrice d'information mutuelle
2. Seuillage de la matrice pour retenir les valeurs significatives et donc les liens significatifs
3. Suppression des liens indirects grâce à la Data Processing Inequality (DPI)

Information mutuelle et Data Processing Inequality (DPI)

Si 2 gènes g_1 et g_3 interagissent seulement à travers l'action d'un troisième gène g_2, alors l'information mutuelle respecte l'inégalité suivante :

$$I(g_1, g_3) \leqslant \min[I(g_1, g_2); I(g_2, g_3)]$$

Dans un triplet connecté, la plus faible des 3 informations mutuelles provient d'un lien indirect, qui sera éliminé.
Suppression des liens indirects dans ARACNE

1. Déterminer tous les triplets connectés (avec une *MI* supérieure au seuil de significativité)
2. Dans chaque triplet, retirer le lien correspondant à la plus faible *MI*

ARACNE / exemple de post-traitement du réseau par la DPI

Figure 2: Réseau initial

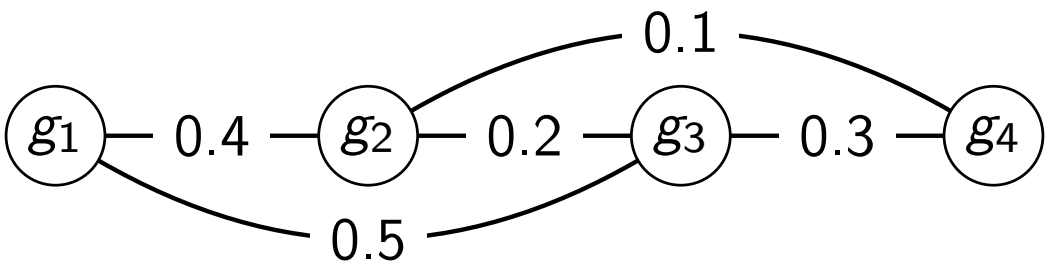
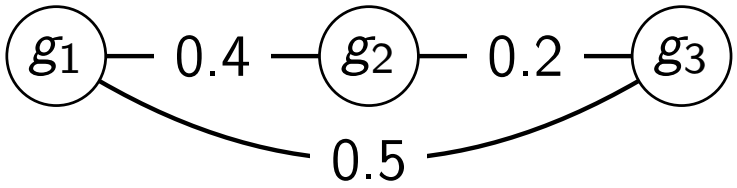
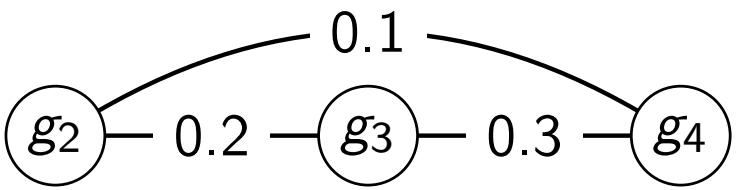
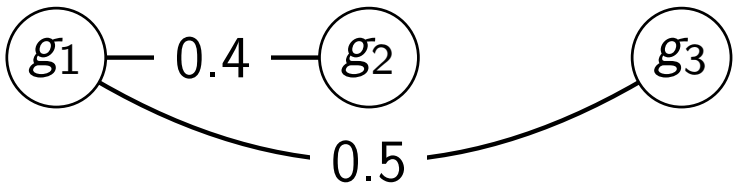


Figure 3: Identification et simplification des triplets.







ARACNE exemple de post-traitement du réseau par la DPI

Figure 4: Réseau initial

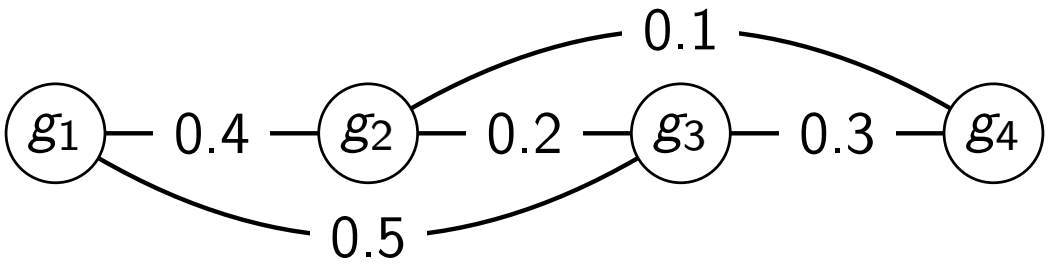
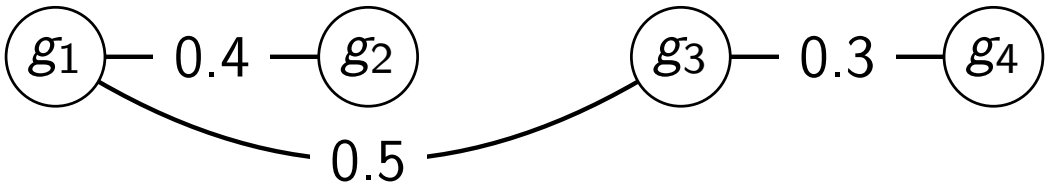


Figure 5: Réseau final



ARACNE Complexité et disponibilité

- Complexité Pour un jeu de données constitué de n gènes et p échantillons
  - l'estimation de l'information mutuelle est en $\mathcal{O}(n^2p^2)$
  - l'utilisation de la DPI est en $\mathcal{O}(n^3)$

  ARACNE a donc une complexité en $\mathcal{O}(n^3 + n^2p^2)$.
- Code: Disponible dans le package minet (Mutual Information NETworks) du projet Bioconductor (R).

Section 4

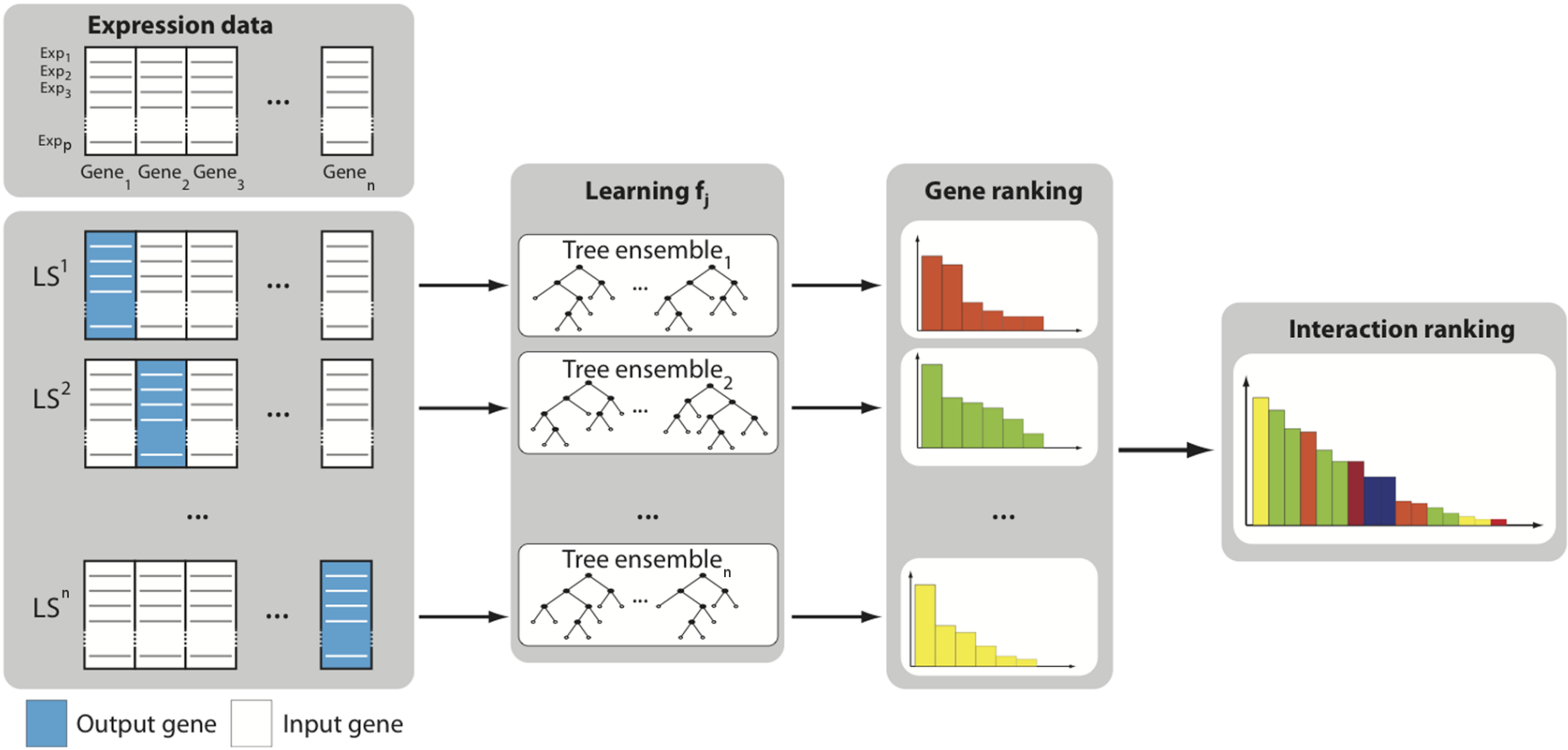
Exemple de méthodes utilisant la régression

- GENIE3
- TIGRESS

Principe Décomposer l'inférence d'un réseau entre n gènes, en n problèmes différents de régression. On cherche à expliquer le vecteur d'expression d'un gène cible à partir des vecteurs d'expression des autres gènes

GENIE3 GEne Network Inference with Ensemble of trees

Huynh-Thu et al. (2010)



GENIE3

- Décompose l'inférence d'un réseau entre n gènes, en n problèmes différents de régression $\{f_j | j = 1, \ldots, i = n\}$.
- Pour chaque problème f_j, l'expression d'un gène g_j est prédite en fonction des autres gènes par une forêt aléatoire.
- Chaque f_j donne un classement de l'importance des gènes g_i dans la prédiction de l'expression de g_j.
- Les différents classements sont agrégés.
  w_{ij} est le poids de g_i pour la prédiction de g_j.
- On obtient donc des liens orientés et pondérés avec cette méthode.

Rappel : arbre de régression (1/2)

- Échantillon de n points observés.
- Pour le point i de l'échantillon, l'expression du gène à prédire est une valeur continue noté y_i.
- Pour ce même point, l'expression des autres gènes (variables explicatives), sont notées x_i^j, avec $j = 1, \ldots, d$.
- Objectif : découper l'espace des variables explicatives en régions R_1, ..., R_J (les feuilles de l'arbre) qui minimisent:

$$RSS = \sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_j)^2, \tag{1}$$

avec, $\hat{y}_j = \frac{1}{n_j} \sum_{i \in R_j} y_i$, où n_j est le nombre d'observations dans la feuille R_j.

Algorithme - arbre de régression

- $R_-(j,s)$ l'ensemble des points tel que $x^j < s$.
- $R_+(j,s)$ l'ensemble des points tel que $x^j \geq s$.
- A chaque étape, pour construire l'arbre de régression, on choisit la la variable j et le seuil s minimisant

$$\frac{1}{n_{R_-}} \sum_{i \in R_-(j,s)} (y_i - \hat{y}_{R_-})^2 + \frac{1}{n_{R_+}} \sum_{i \in R_+(j,s)} (y_i - \hat{y}_{R_+})^2 \qquad (2)$$

- $Var(R_-) = \frac{1}{n_{R_-}} \sum_{i \in R_-(j,s)} (y_i - \hat{y}_{R_-})^2$ est la variance intra-groupe associée à l'ensemble de points $R_-(j,s)$.
- $Var(R_+) = \frac{1}{n_{R_+}} \sum_{i \in R_+(j,s)} (y_i - \hat{y}_{R_+})^2$ est la variance intra-groupe associée à l'ensemble de points $R_+(j,s)$.

GENIE3. Importance d'un lien

- Dans un arbre de régression, l'importance d'un noeud $\mathcal{N}$ est mesurée par la réduction de variance due à ce test

$$I(\mathcal{N}) = n_S Var(S) - n_{R_-} Var(R_-) - n_{R_+} Var(R_+) \tag{3}$$

  où S est l'ensemble d'exemples de ce noeud, R_- et R_+ les 2 ensembles résultant du split selon une variable explicative donnée (cf. équation 2).
- L'importance d'une variable dans un arbre est la somme des valeurs $I(\mathcal{N})$ pour tous les noeuds ($\mathcal{N}$) où cette variable est utilisée.
- Pour une forêt aléatoire, l'importance d'une variable est la moyenne des valeurs obtenues sur les différents arbres.

GENIE3. Importance d'un lien

- Chaque problème de régression f_j donne donc un ensemble de poids w_{ij} qui mesurent l'importance du gène i dans la prédiction du gène j.
- Pour se servir de ces mesures pour ordonner tous les liens on doit normaliser les expressions de gènes afin qu'ils aient tous une variance de 1 dans l'ensemble initial.

GENIE3. Paramètres de la méthode

- Pour chaque problème f_j, 1000 arbres sont construits
- K, le nombre d'attributs choisis aléatoirement pour chaque split-test de l'arbre, est fixé à $\sqrt{n-1}$

GENIE3 Complexité et disponibilité

- Complexité :
  Pour un jeu de données constitué de n gènes et p échantillons
  T le nombre d'arbres construits pour chaque forêt
  K le nombre d'attributs utilisés
  La complexité de GENIE3 est en $\mathcal{O}(nTKp \log p)$
- Implémentations en Mathlab, R et Python disponibles sur:
  https://github.com/vahuynh/GENIE3
  et aussi dans Bioconductor

TIGRESS: Trustful Inference of Gene REgulation using Stability Selection

Haury et al. (2012)

- Comme GENIE3, traite l'inférence d'un réseau à travers plusieurs problèmes de régression.
- Pour chaque problème f_j, l'expression d'un gène g_j est traité comme un problème de régression linéaire avec sélection parcimonieuse: méthode LARS (Least Angle Regression)
- Introduit une méthode de stabilité pour la sélection qui permet d'agréger les scores issus de chaque régression et affecte à un score pour chaque lien du réseau

Méthode LARS

Efron et al. (2004)

- Comme pour Lasso, modèle linéaire avec bruit additif Gaussien :

$$y := \beta_1 x_1 + \cdots + \beta_d x_d + \epsilon \tag{4}$$

- Au lieu d'utiliser une pénalisation ℓ_1 sur les coefficient β_j, l'idée est de partir d'un modèle avec aucune variables sélectionnées, puis introduire ces variables les une après les autres jusqu'à ce que tout le signal soit reconstruit.
- Avantage : pas besoin de spécifier d'hyperparamètre λ qui régule le niveau pénalisation ℓ_1.

Etapes de l'algorithme LARS

1. Commencer avec tous les coefficients β_j égaux à zéro.
2. Trouver la variable explicative x_j la plus corrélée avec y.
3. Augmenter le coefficient β_j dans la direction du signe de sa corrélation avec y. Calculer les résidus $r = y - \hat{y}$ au fur et à mesure. On s'arrête lorsqu'un un autre prédicteur x_k a autant de corrélation avec r que x_j en avait avec y.
4. Augmenter les coefficients (β_j, β_k), jusqu'à ce qu'un autre prédicteur x_m ait autant de corrélation avec le résidu r.
5. Continuer jusqu'à ce qu'un critère d'arrêt soit atteint.
6. L'ordre de sélection des variables explicatives donne une mesure de leur importance pour reconstruire la variable cible y.

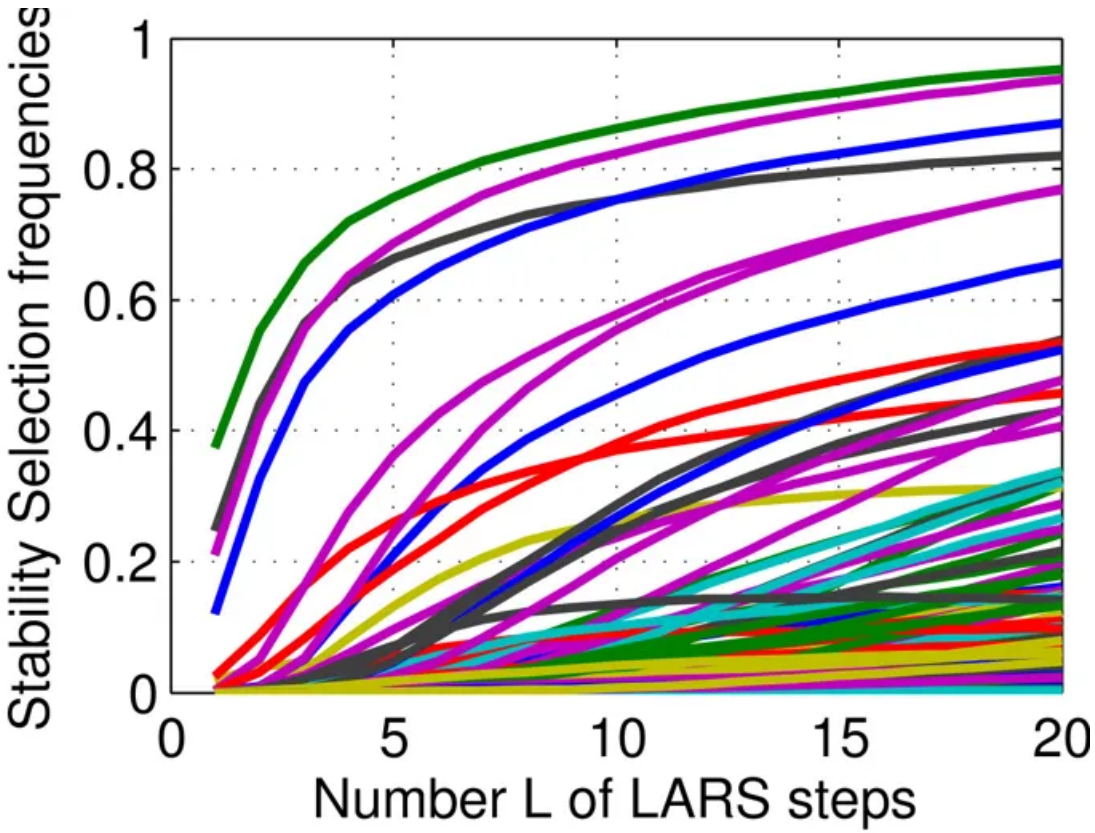
TIGRESS - LARS et stability selection

- **Sélection parcimonieuse**
  - LARS : introduction pas à pas des variables
  - après L étapes dans LARS, L variables sont sélectionnées
  - mais sensibles aux fortes corrélations entre variables
- **Stabilité de la sélection**
  - Recommencer un grand nombre de fois le processus de sélection sur des données perturbées.
  - La fréquence de sélection d'un attribut donne un score de pertinence pour cet attribut.

Stability selection

- Exécuter R fois LARS (R=1000 par exemple) sur un jeu d'apprentissage obtenu par
  - ré-échantillonage des expériences
  - Perturbation des variables explicatives : multiplier chaque variable explicative par un nombre aléatoire tiré uniformément dans $[\alpha, 1]$
  - Pour chaque run de LARS, on a après L pas de sélection une liste triée de L facteurs explicatifs
- $F(g, t, l)$: fréquence de sélection du facteur t parmi les l premiers facteurs dans la prédiction de g, pour $g \in \mathcal{G}, t \in \mathcal{T}, l \in [1, L]$

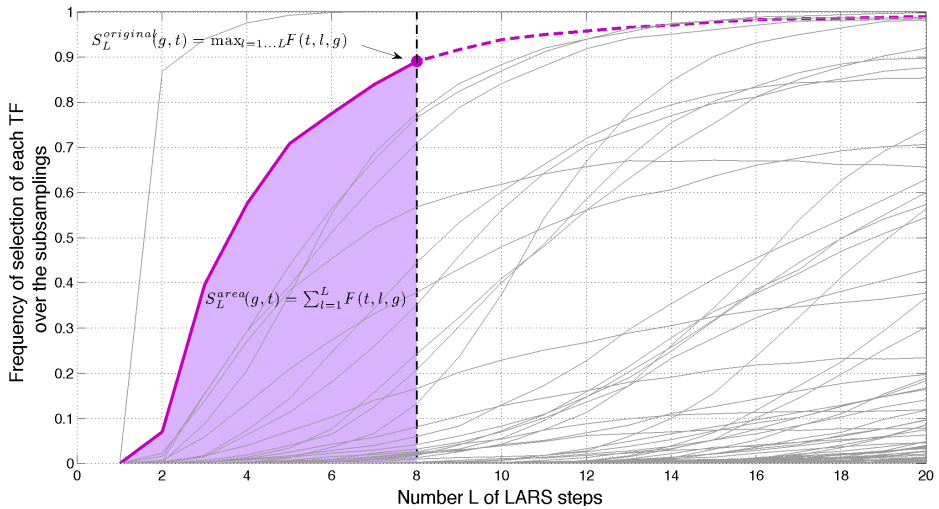
Stability selection



Pour un gène $g \in \mathcal{G}$, courbes de fréquences des différents facteurs de transcription

Stability selection

On doit agréger les $F(g, t, l)$ en un unique score $s(g, t)$ pour chaque lien candidat $s(g, t)$.



Pour un L donné au lieu de considérer la valeur max de F, on considère l'aire sous la courbe de F: prend en compte la position de t parmi les premiers

TIGRESS - Paramètres

- R: nombre de runs de LARS (aussi grand que possible)
- L: nombre d'étapes de sélection dans LARS
- α: randomization des valeurs d'expression
- K: le nombre de liens retenus dans le réseau (Les K meilleures valeurs du score $s(g, t)$)

TIGRESS - Complexité et disponibilité

- Complexité
  Pour un jeu de données constitué de n gènes et p échantillons, avec q facteurs de transcription
  un run de LARS avec L étapes est en $\mathcal{O}(pqL + L^3)$
  R runs pour chacun des n gènes donc complexité de TIGRESS en $\mathcal{O}(nR(pqL + L^3))$
- Implémentation en Mathlab, disponible à:
  http://members.cbio.mines-paristech.fr/ ahaury/svn/dream5/html/index.html

Section 5

Données synthétiques et réelles pour l'inférence de réseaux

Validation des méthodes d'inférence de graphes sur des données synthétiques

1. Generation of DAG with 20 and 100 variables.
2. DAG structure such that the number of parents for each variable is uniformly drawn in $\{0, \ldots, 5\}$;
3. For the i-th DAG, the mean μ_i and variance σ_i of the noise variables are drawn as $\mu_i \sim \mathbb{U}(-2, 2)$ and $\sigma_i \sim \mathbb{U}(0, 0.4)$ and the distribution of the noise variables is set to $\mathcal{N}(\mu_i, \sigma_i)$;
4. For each graph, a 500 samples data set is iid generated following the topological order of the graph, with for $\ell = 1$ to 500,

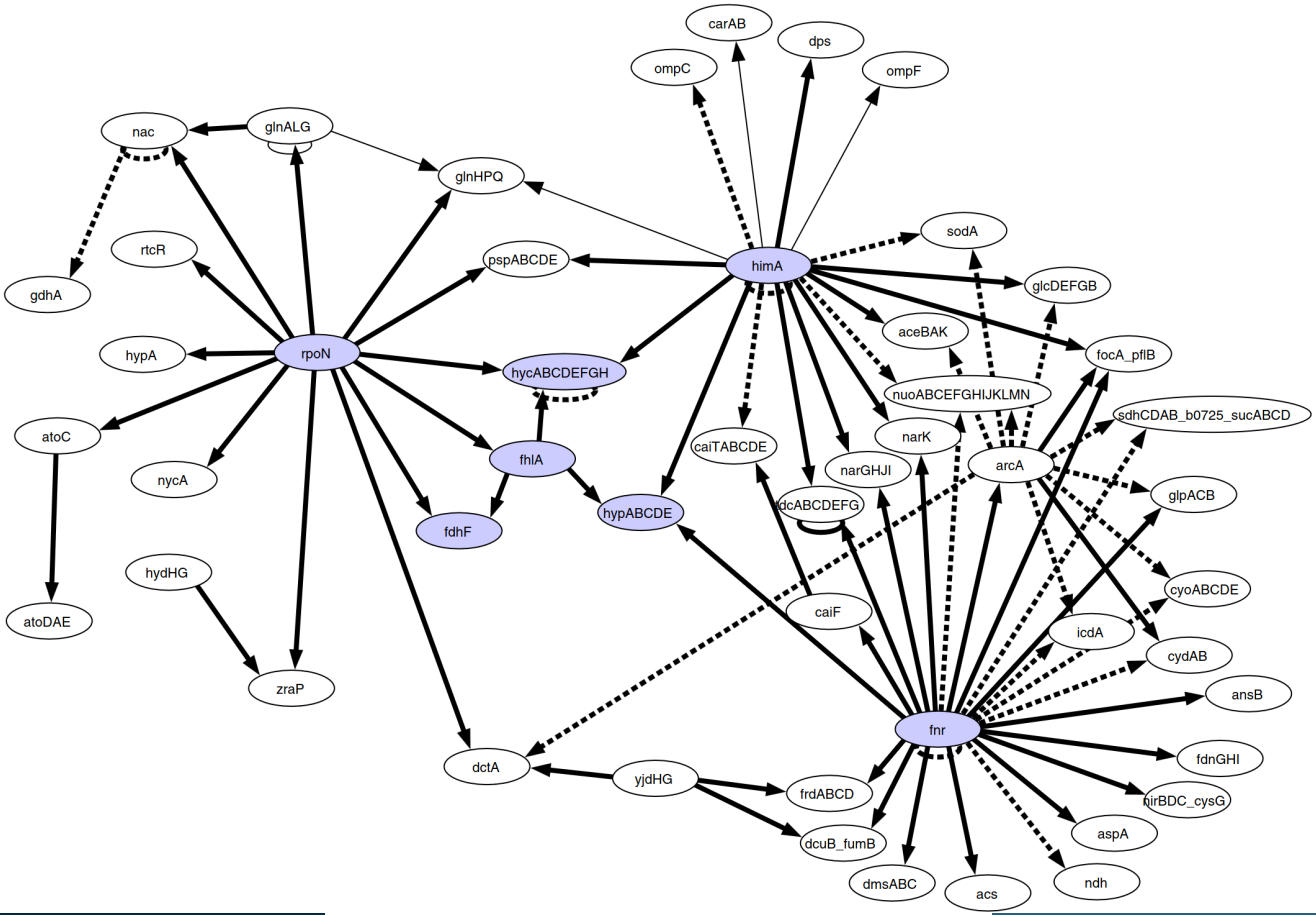
$$x^{(\ell)} = (x_1^{(\ell)}, \ldots, x_d^{(\ell)}), \quad x_i^{(\ell)} \sim f_i(X_{\mathsf{Pa}(i)}, E_i), \text{ with } E_i \sim \mathcal{N}(\mu_i, \sigma_i).$$

Six categories of causal mechanisms

1. *Linear*: $X_i = \sum_{j \in \mathrm{Pa}(i)} a_{i,j} X_j + E_i$
2. *Sigmoid AM*: $X_i = \sum_{j \in \mathrm{Pa}(i)} f_{i,j}(X_j) + E_i$, where $f_{i,j}(x_j) = a \cdot \frac{b \cdot (x_j + c)}{1 + |b \cdot (x_j + c)|}$ with $a \sim \mathrm{Exp}(4) + 1$, $b \sim \mathcal{U}([-2, -0.5] \cup [0.5, 2])$ and $c \sim \mathcal{U}([-2, 2])$.
3. *Sigmoid Mix*: $X_i = f_i(\sum_{j \in \mathrm{Pa}(i)} X_j + E_i)$, where f_i is as in the previous bullet-point.
4. *GP AM*: $X_i = \sum_{j \in \mathrm{Pa}(i)} f_{i,j}(X_j) + E_i$ where $f_{i,j}$ is an univariate Gaussian process with a Gaussian kernel of unit bandwidth.
5. *GP Mix*: $X_i = f_i([X_{\mathrm{Pa}(i)}, E_i])$, where f_i is a multivariate Gaussian process with a Gaussian kernel of unit bandwidth.
6. *NN*: $X_i = f_i(X_{\mathrm{Pa}(i)}, E_i)$, where f_i a random single-hidden-layer neural network with 20 ReLU units.

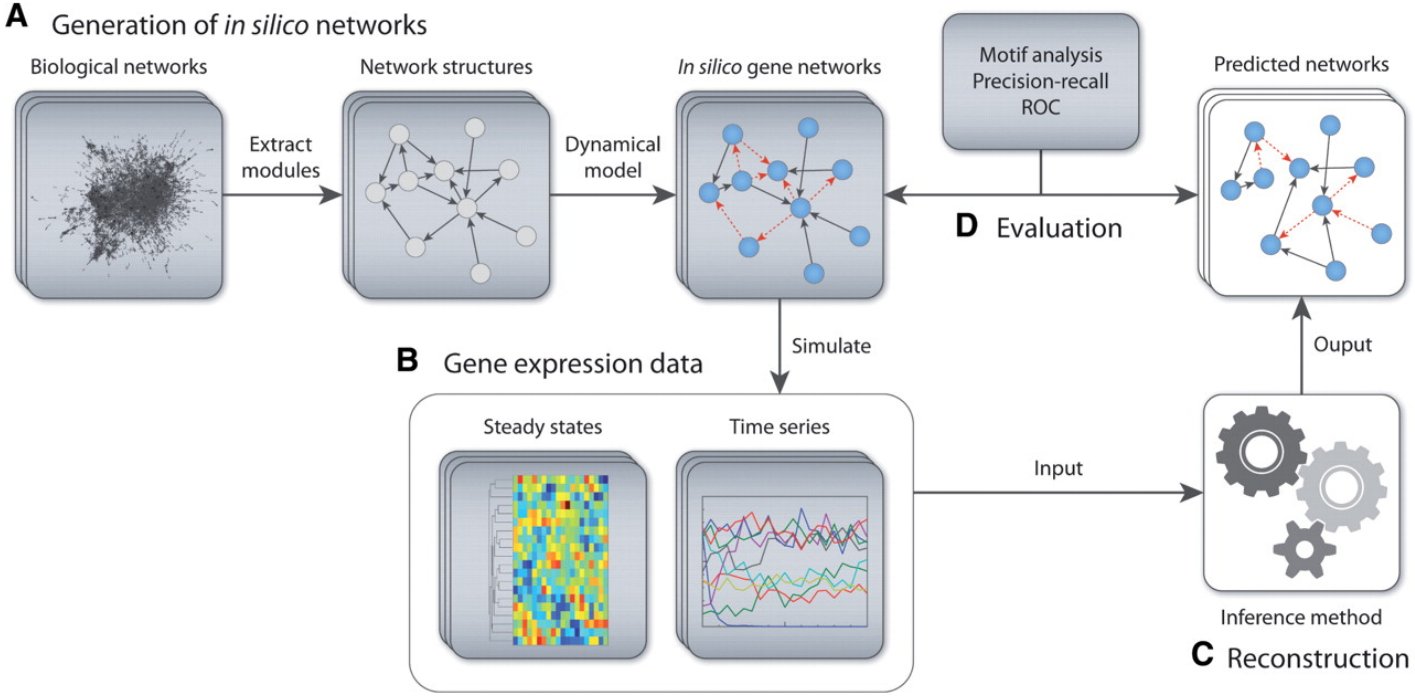
Validation des méthodes d'inférence sur des données synthétiques issues de simulateurs biologique réalistes

SynTReN: a generator of synthetic gene expression data for design and analysis of structure learning algorithms (2006).



Un autre générateur de données biologiques artificielles

- Challenges Dream 4 et Dream 5
- GeneNetWeaver Schaffter et al. (2011)



Compétitions Dream5 (http://dreamchallenges.org)

- *Network 1* is a simulated network with simulated expression data (GeneNetWeaver software).
- Other expression data sets are real expression data collected for E. coli (*Network 3*) and S. cerevisiae (*Network 4*).
- On these data sets, the set $\mathcal{T}$ of potential causes (Transcription Factors or TF) is known and constitutes a subset of the genes ($\mathcal{T} \subset \mathcal{G}$).
- The task is to infer all directed edges (t, g) with $t \in \mathcal{T}$ and $g \in \mathcal{G}$.

Network	# TF	# Genes	# Observations	# Verified interactions
DREAM5 Network 1 (in-silico)	195	1643	805	4012
DREAM5 Network 3 (E.coli)	334	4511	805	2066
DREAM5 Network 4 (S.cerevisiae)	333	5950	536	3940

Table 1: Dream5 challenge

Validation des méthodes d'inférence sur des données réelles

Gene expression data including 7,466 observational samples for 11 proteins (variables).

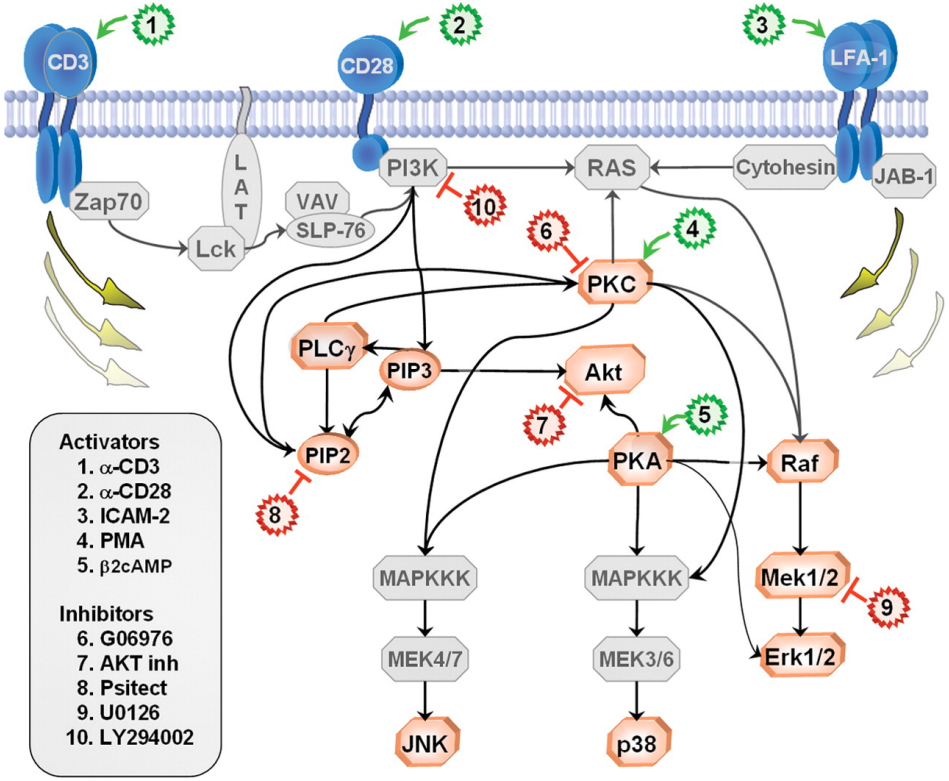
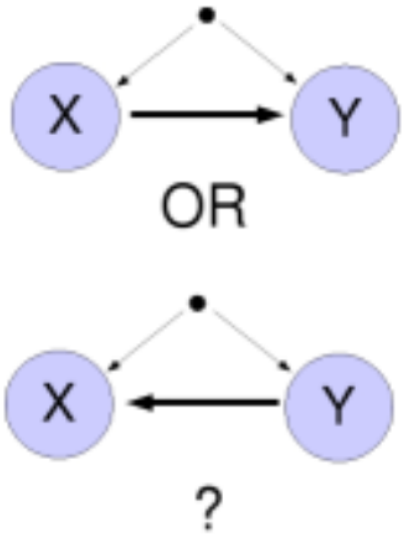


Figure 6: Conventionally accepted signaling molecule interactions between the 11 variables of the data set.

Données réelles pour la causalité paire à paire



https://webdav.tuebingen.mpg.de/cause-effect/

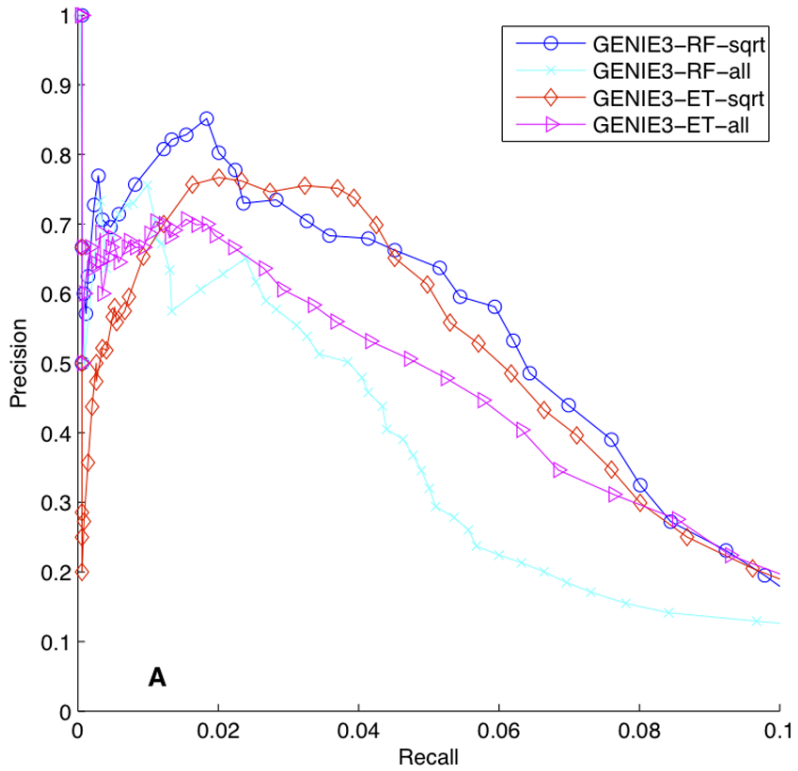
Section 6

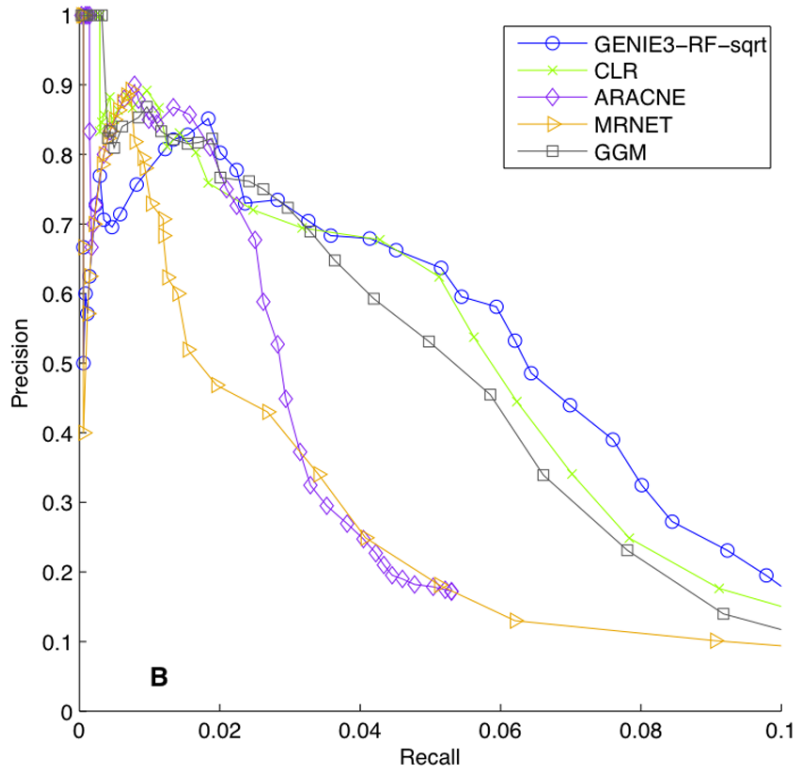
Indicateurs de performance pour la causalité

Area under the Precision Recall Curve (AUPR)

- Causation score $c_{i,j}$ given by a method for each edge $X_i \to X_j$ in $\widehat{\mathcal{G}}$.
- True DAG $\mathcal{G}$.
- A true positive is an edge $X_i \to X_j$ of the true DAG $\mathcal{G}$ which is correctly recovered by the algorithm.
- A false negative is an edge of $\mathcal{G}$ which is missing in$\widehat{\mathcal{G}}$; F_n is the number of false negatives.
- A false positive is an edge in$\widehat{\mathcal{G}}$ which is not in $\mathcal{G}$.
- F_p is the number of false positives.
- Precision-recall curve, showing the tradeoff between precision ($T_p/(T_p + F_p)$) and recall ($T_p/(T_p + F_n)$) for different causation thresholds.

Exemple : courbe de précision-rappel - méthode GENIE3





E. Coli

Courbes précision-rappel pour différents paramétrages de Genie3 et pour Genie3 vs d'autres méthodes de l'état de l'art

Structural Hamming Distance (SHD)

- Corresponds to number of missing edges and redundant edges in the found structure.
- SHD score is computed by considering all edges $X_i \to X_j$ in $\widehat{\mathcal{G}}$ recovered by a given method with $c_{i,j} > .5$
- A reversal error (retaining $X_j \to X_i$ while $\mathcal{G}$ includes edge $X_i \to X_j$) is counted as a single mistake:

$$\text{SHD}(\hat{A}, A) = \sum_{i,j} |\hat{A}_{i,j} - A_{i,j}| - \frac{1}{2} \sum_{i,j} (1 - max(1, \hat{A}_{i,j} + A_{j,i})), \quad (5)$$

with A (respectively $\hat{A}$) the adjacency matrix of $\mathcal{G}$ (resp. the found causal graph $\widehat{\mathcal{G}}$).

Section 7

Quelques résultats empiriques sur l'inférence de réseaux

Résultats sur des graphes synthétiques de 20 variables

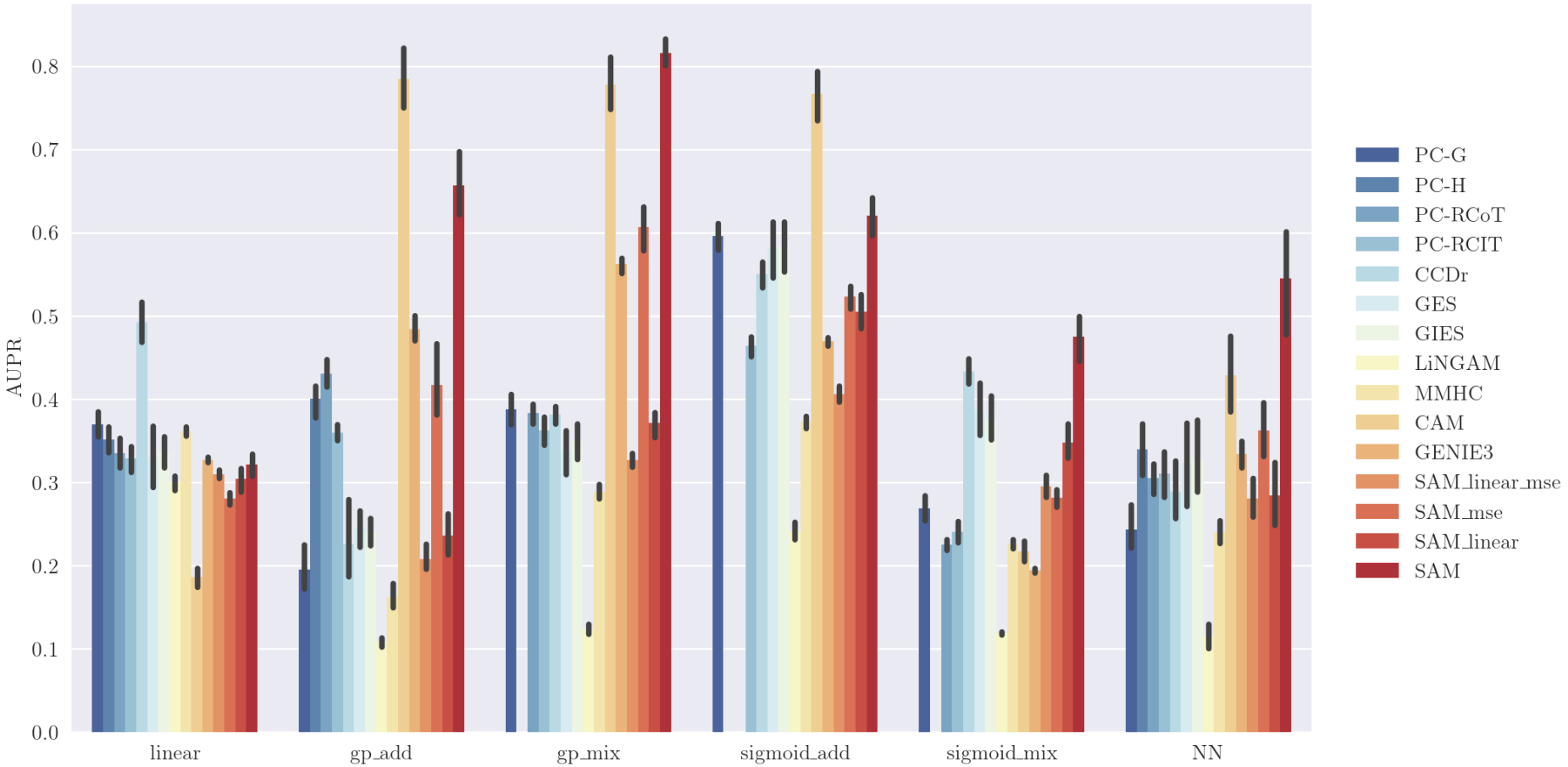


Figure 7: AUPR Scores on 20-node synthetic graphs; the error bar indicates the standard deviation.

Résultats sur des graphes synthétiques de 100 variables

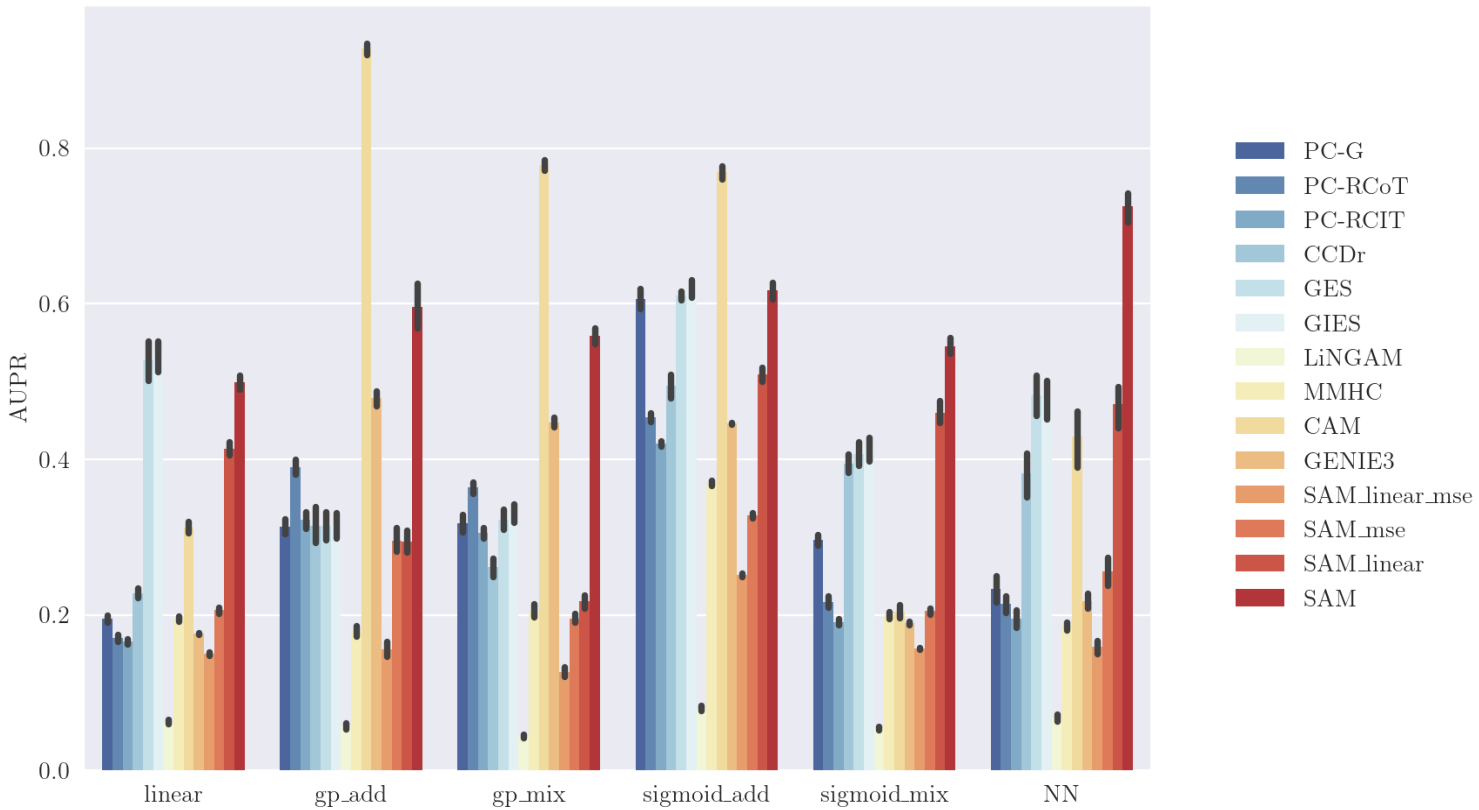


Figure 8: AUPR Scores on 100-node synthetic graphs; the error bar indicates the standard deviation.

SynTREN Simulator - Scores

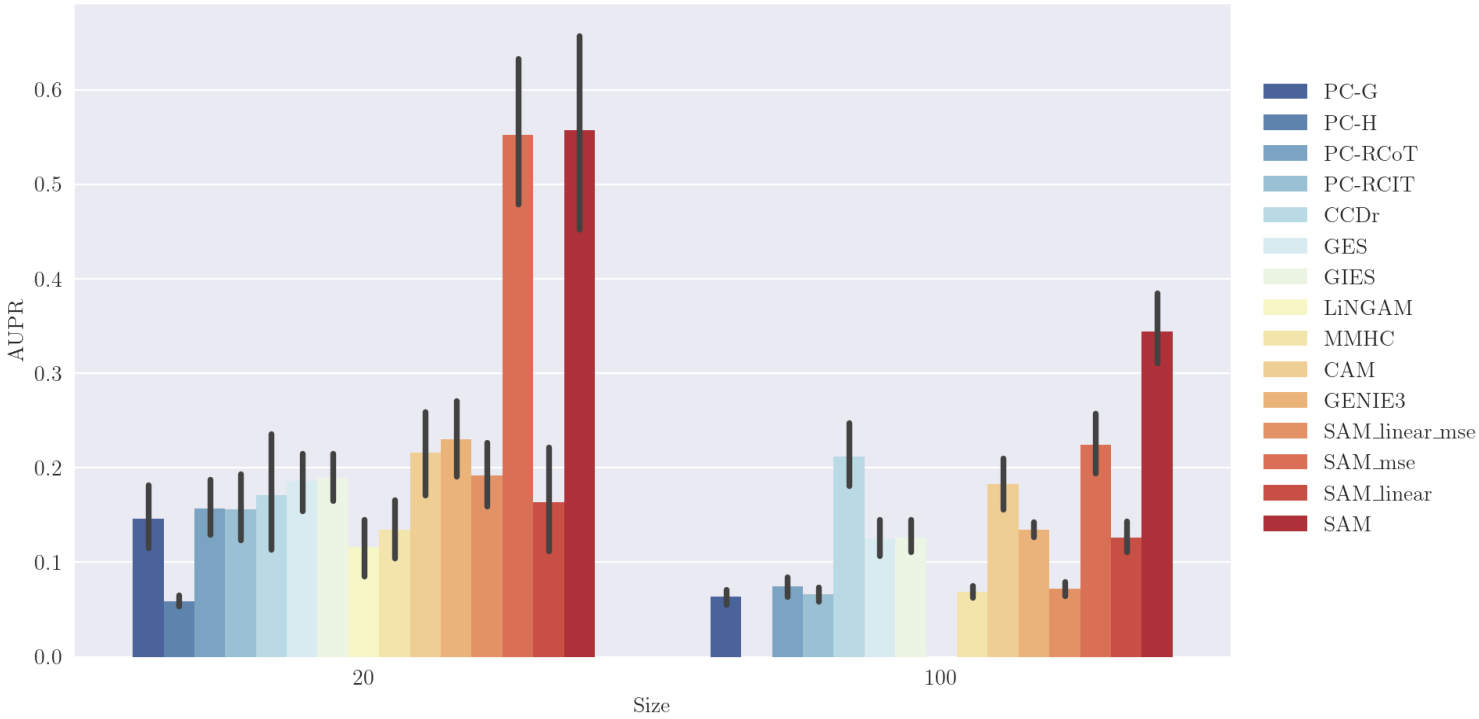


Figure 9: AUPR Scores on SynTREN graphs; the Figure bar indicates the standard deviation. Left: 20 nodes. Right: 100 nodes.

.

SynTREN Precision Recall curve

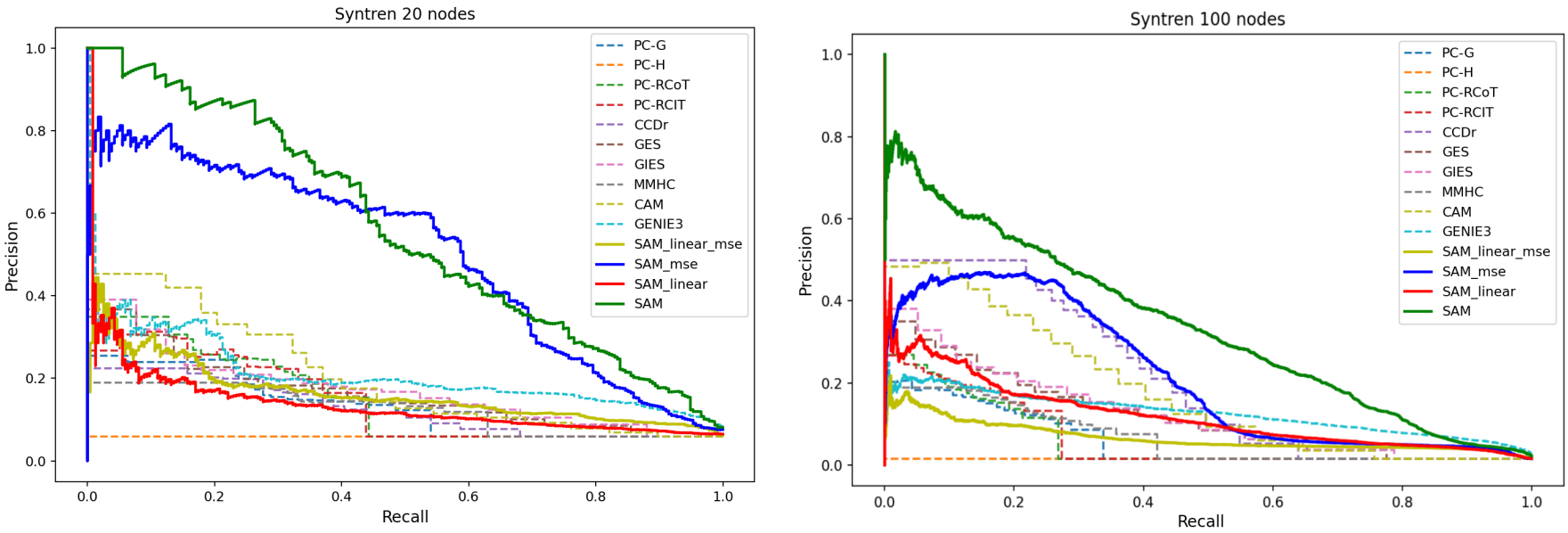


Figure 10: Precision/Recall curve for SynTREN graphs: Left, 20 nodes; Right, 100 nodes.

GeneNetWeaver Simulator - DREAM4

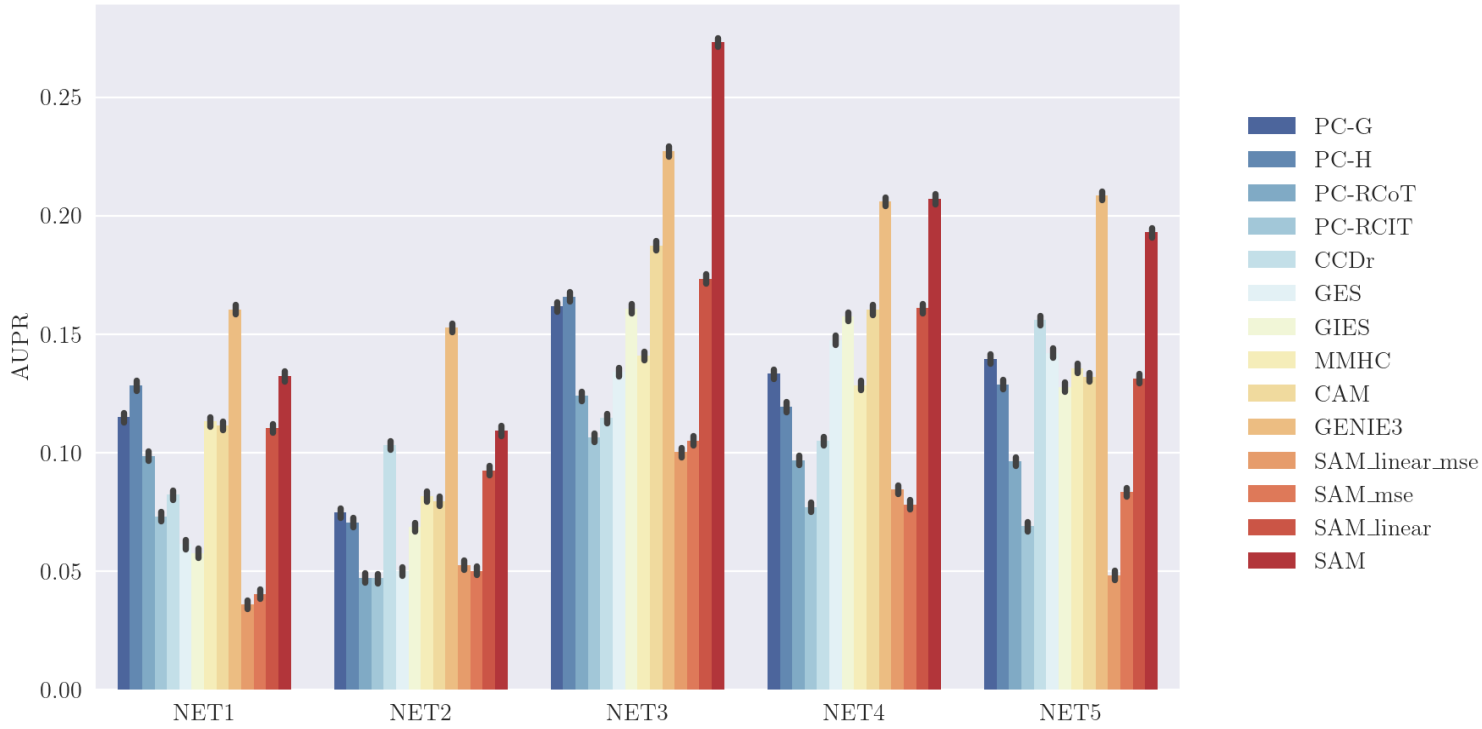


Figure 11: Precision/Recall curve for the Dream4 *In Silico Multifactorial Challenge*

GeneNetWeaver Simulator - DREAM5

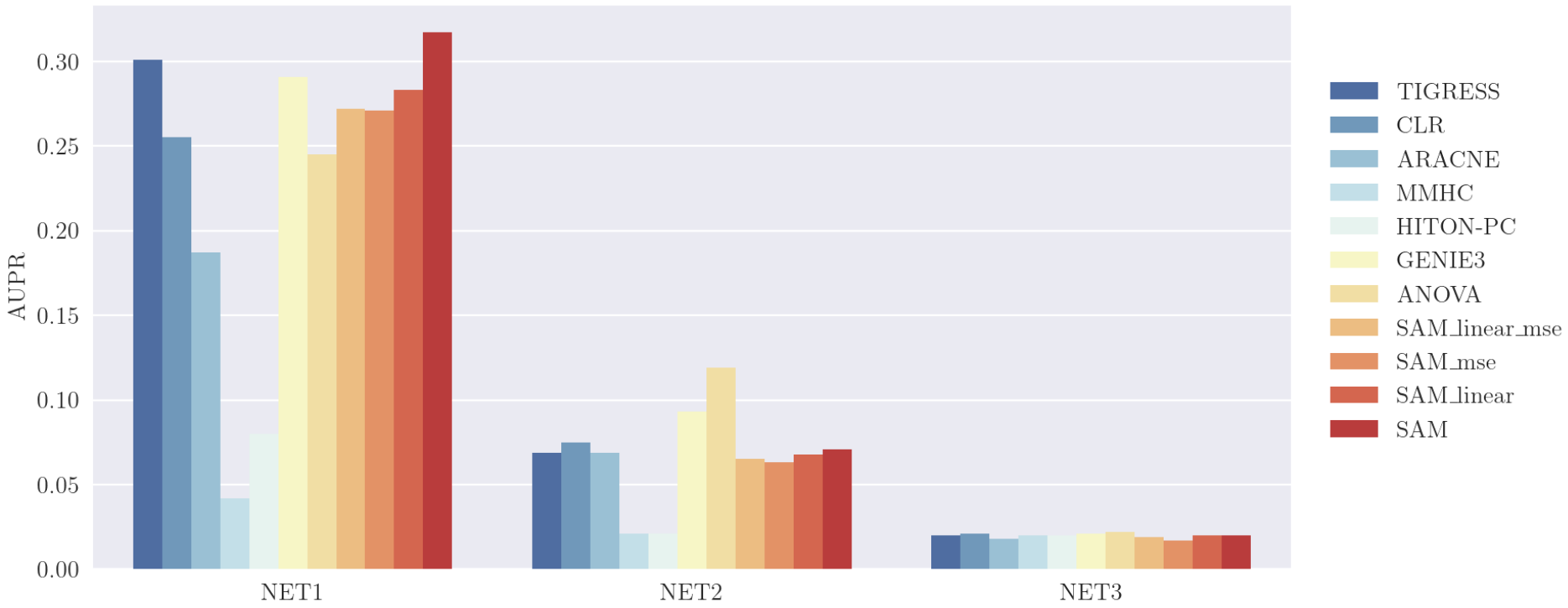


Figure 12: Performance of causal graph discovery methods on the three networks of the Dream5 Challenge.

Section 8

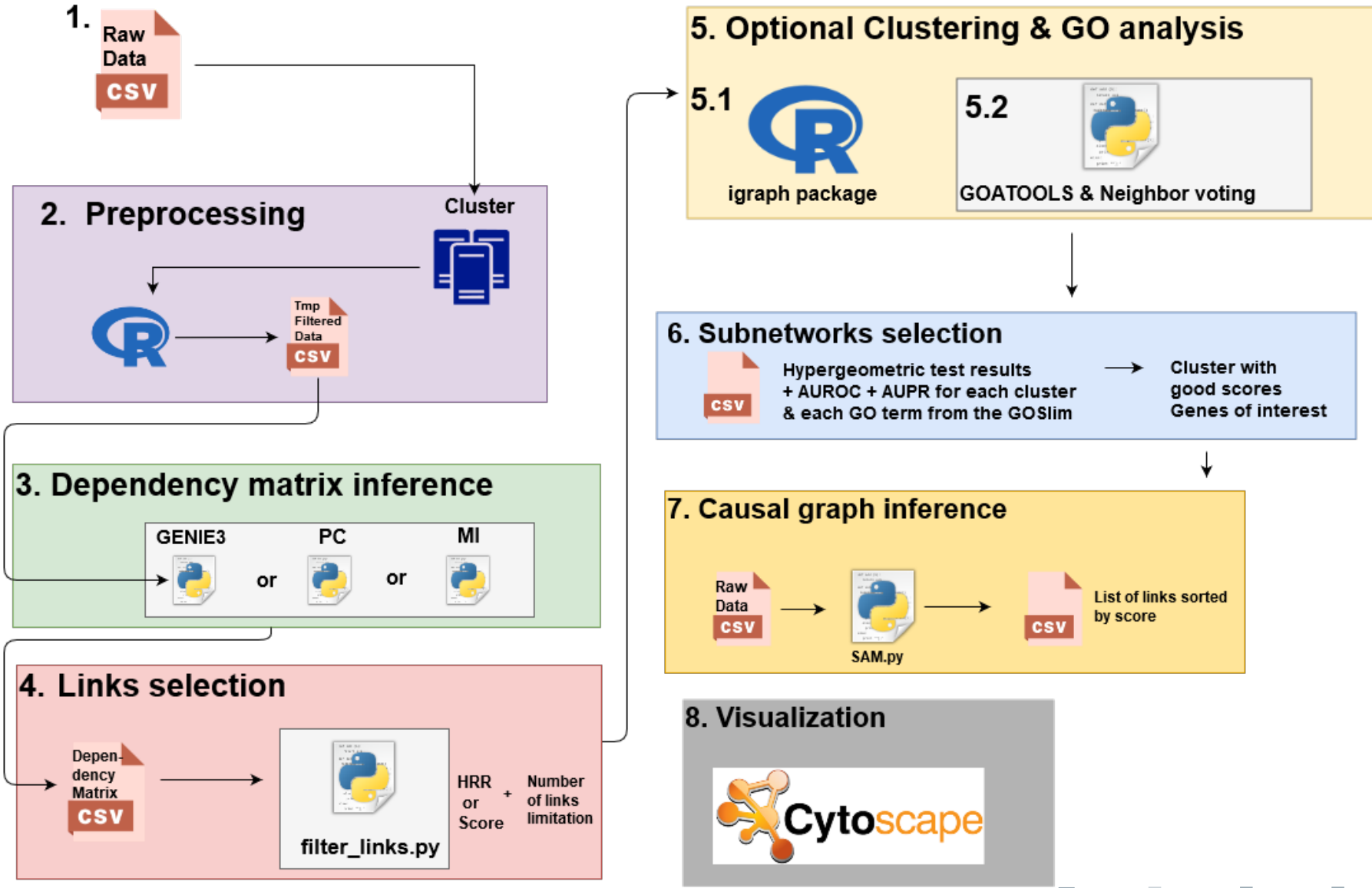
L'inférence de réseaux en pratique

Librairies pour l'inférence de réseaux

- Causal Inference in R. `https://www.r-causal.org/`.
- Causal inference in Python.

`https://fentechsolutions.github.io/CausalDiscoveryToolbox/html/index.html`.

Un exemple de workflow pour traiter un problème biologique



Bibliographie

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression.

Farhadian, M., Rafat, S. A., Panahi, B., and Mayack, C. (2021). Weighted gene co-expression network analysis identifies modules and functionally enriched pathways in the lactation process. *Scientific Reports*, 11(1):1–15.

Haury, A.-C., Mordelet, F., Vera-Licona, P., and Vert, J.-P. (2012). TIGRESS: Trustful Inference of Gene REgulation using Stability Selection. *BMC Systems Biology*, 6:145.

Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., and Geurts, P. (2010). Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. *PLoS ONE*, 5(9):e12776.

Langfelder, P. and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, 9(1):559.

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R. D., and Califano, A. (2006). ARACNE: An Algorithm for the