

Méthodes statistiques pour l'inférence de réseaux de régulation

Océane Cassan, Sophie Lèbre

oceane.cassan@cnrs.fr

sophie.lebre@umontpellier.fr

BFP M1, Parcours Bipa

January 24, 2022

① Introduction

② Méthodes classiques pour l'inférence de GRN

③ Inférence de GRN par régression

④ 2 exemples de méthodes par régression

⑤ Validation et perspectives

Introduction

Exemple de réseau de régulation

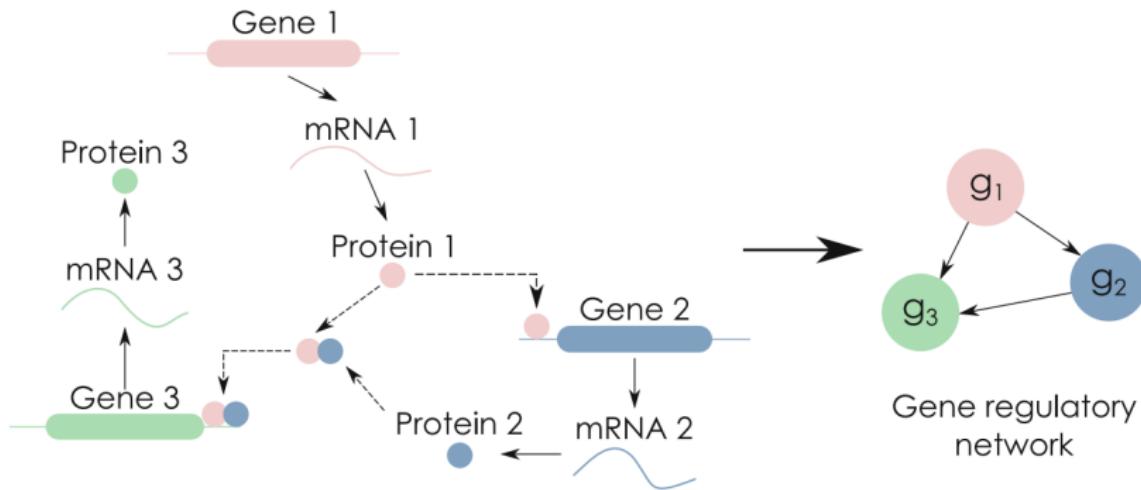
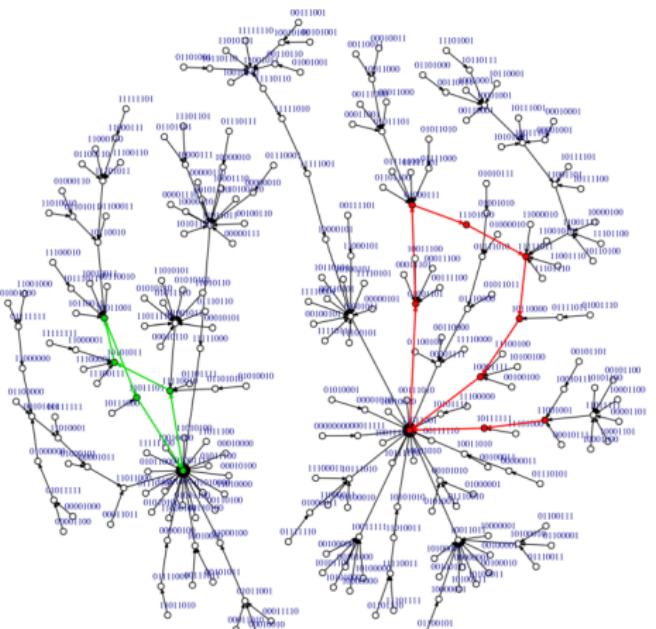


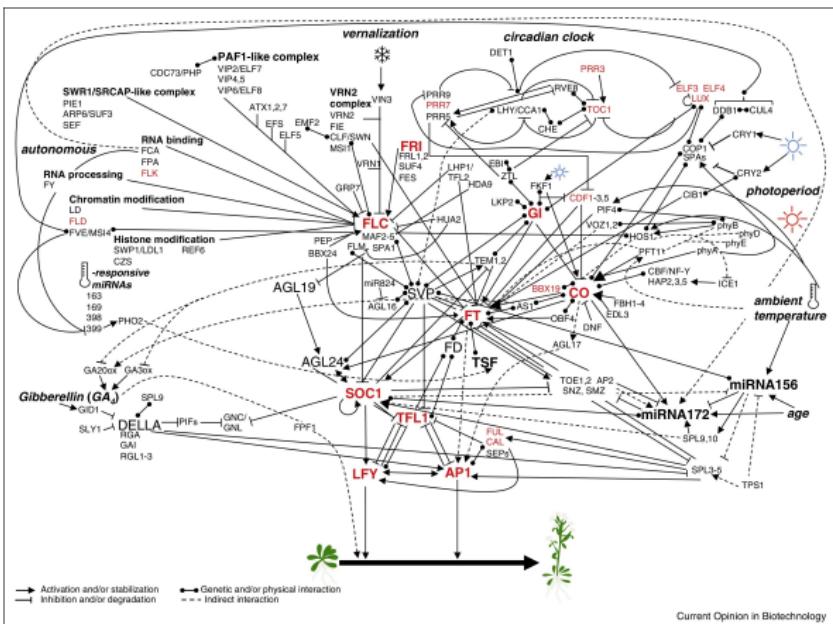
Figure: Illustration schématique d'un réseau de régulation entre 3 gènes [Sanguinetti and Huynh-Thu, 2019]

Exemple de réseau de régulation chez Arabidopsis



Timmermann, T., González, B. & Ruz, G.A. Reconstruction of a gene regulatory network of the induced systemic resistance defense response in *Arabidopsis* using boolean networks. BMC Bioinformatics 21, 142 (2020). A set of small, light-blue navigation icons typically used in Beamer presentations for navigating between slides and sections.

Exemple de réseau de régulation chez Arabidopsis



Flowering time gene network with known genetic and epigenetic regulators in *Arabidopsis thaliana*.

Blümel, et al., 2015, Current Opinion in Biotechnology

Utilisation des réseaux en génomique fonctionnelle



Différents domaines d'application :

- Biologie fondamentale
 - Fonction et interaction des gènes et protéines (annotations, base de données), Machinerie cellulaire
- Santé, Médecine
 - Diagnostique (ex: type de tumeur cancéreuse), Pronostique (ex: analyse de survie), Thérapie (ex: prédiction de la réponse à un traitement, médecine personnalisée)
- Biologie végétale
 - Agriculture, alimentation, adaptation au changement climatique

Utilisation des réseaux en génomique fonctionnelle



Différents niveaux d'observation :

- **Génome** : analyse de séquence (ex: détection de motifs)
- **Transcriptome** : quantité d'ARN messager
- **Protéome** : quantité de protéines / interactions et fonctions des protéines

Données expérimentales pour l'inférence d'un GRN

- Les **données de séquences génomiques** (séquence ADN du génome, annotations, motifs de fixation des TF, PWM, ...)

Données expérimentales pour l'inférence d'un GRN

- Les **données de séquences génomiques** (séquence ADN du génome, annotations, motifs de fixation des TF, PWM, ...)
- Les **données d'expression (micro array et RNA-Seq)**
 - Permet de quantifier le niveau d'expression (nombre de transcrits) pour **l'ensemble des gènes d'un organisme** dans une condition donnée.
 - L'expression est une conséquence indirecte, il faut donc "démêler", "deviner", ce qui est de la régulation, ou ce qui est de la co-expression fortuite ou induite par des régulations communes

Données expérimentales pour l'inférence d'un GRN

- Les **données de séquences génomiques** (séquence ADN du génome, annotations, motifs de fixation des TF, PWM, ...)
- Les **données d'expression (micro array et RNA-Seq)**
 - Permet de quantifier le niveau d'expression (nombre de transcrits) pour **l'ensemble des gènes d'un organisme** dans une condition donnée.
 - L'expression est une conséquence indirecte, il faut donc "démêler", "deviner", ce qui est de la régulation, ou ce qui est de la co-expression fortuite ou induite par des régulations communes
- Les **données d'accessibilité de la chromatine (ATAC-Seq), données de contacts de la chromatine (HiC-Seq)**



Données d'expression RNA-Seq (genome wide)

Typical gene expression data

regulators

	C_1	C_2	C_3	S_1	S_2	S_3	M_1	M_2	M_3	H_1	H_2	H_3	SM_1	SM_2	SM_3	SH_1	SH_2	SH_3
AT1G01010.1	127	67.9	65.5	94	88.1	95.9	65.1	100.3	126.8	95.4	135	117.2	96.7	104.4	98.1	94.7	96.1	101.3
AT1G01020.1	207.9	220.8	186.8	192.5	225.1	197.8	234.2	196.9	179.4	312.9	366	318	169	179.6	186.5	340.8	352.6	345
AT1G01030.1	32.7	34.4	55.8	33.6	15.9	31.3	21.4	29.8	33.5	47.7	39	51.1	25.2	31.2	34.4	61	65.5	61.2
AT1G01040.2	859.3	978.8	988.6	897.6	837.9	948.6	948.8	903.2	990.1	798.3	701.1	902.9	894.6	880.3	837.2	782.6	776.4	698
AT1G01050.1	846.8	840.2	837.2	836.1	818.8	792.2	802	761.4	719.9	922.2	969.3	930.8	866.9	805.2	840.1	895.2	783.2	849.5
AT1G01060.1	658.2	698.8	636.5	532.7	411	519.7	741.6	782.2	877.7	36.7	20.8	20.5	857.6	853	878.4	29.5	39.6	19.1
• • •																		
AT1G01010.1	127	67.9	65.5	94	88.1	95.9	65.1	100.3	126.8	95.4	135	117.2	96.7	104.4	98.1	94.7	96.1	101.3
AT1G01020.1	207.9	220.8	186.8	192.5	225.1	197.8	234.2	196.9	179.4	312.9	366	318	169	179.6	186.5	340.8	352.6	345
AT1G01030.1	32.7	34.4	55.8	33.6	15.9	31.3	21.4	29.8	33.5	47.7	39	51.1	25.2	31.2	34.4	61	65.5	61.2
AT1G01040.2	859.3	978.8	988.6	897.6	837.9	948.6	948.8	903.2	990.1	798.3	701.1	902.9	894.6	880.3	837.2	782.6	776.4	698
AT1G01050.1	846.8	840.2	837.2	836.1	818.8	792.2	802	761.4	719.9	922.2	969.3	930.8	866.9	805.2	840.1	895.2	783.2	849.5
AT1G01060.1	658.2	698.8	636.5	532.7	411	519.7	741.6	782.2	877.7	36.7	20.8	20.5	857.6	853	878.4	29.5	39.6	19.1

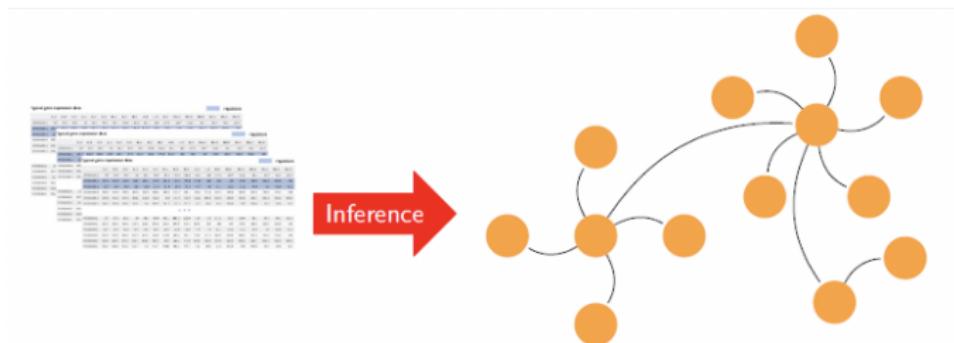
Données expérimentales pour l'inférence d'un GRN

- Les **données de fixation des protéines sur les promoteurs cibles (ChIP-seq)**.

Ces données sont très utiles, cependant :

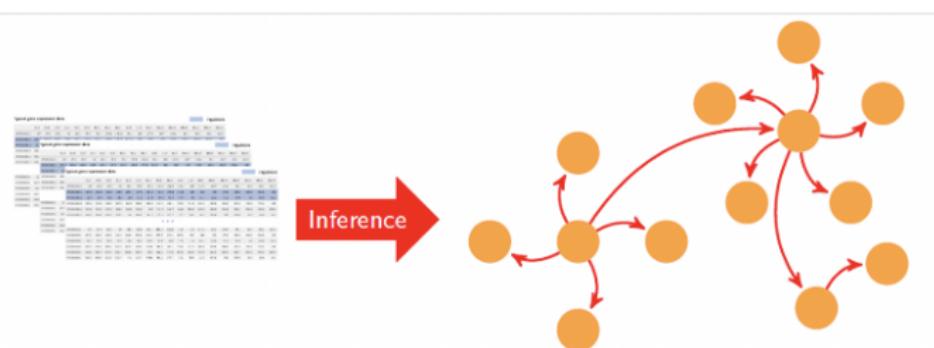
- Leur génération est relativement chère, contraignante donc faisable sur un **nombre restreint de régulateurs**.
- La fixation d'un TF n'implique pas régulation, et la régulation n'implique pas la fixation. (Interactions distantes, transientes, coopération...)

1ère étape : la modélisation



- Que représentent les noeuds ? les arêtes ?
- Biologiquement ? Statistiquement ?
- Le réseau est-il statique ? dynamique ?
- Comment les données ont-elles été générées ? (temporelles (time course), différentes conditions expérimentales (steady state), répliquants)
- Est-ce que les arêtes sont orientées ? (causalité)

1ère étape : la modélisation



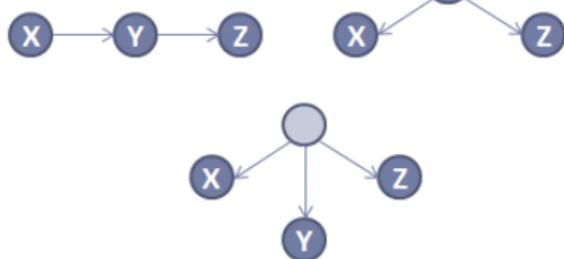
- Que représentent les noeuds ? les arêtes ?
- Biologiquement ? Statistiquement ?
- Le réseau est-il statique ? dynamique ?
- Comment les données ont-elles été générées ? (temporelles (time course), différentes conditions expérimentales (steady state), répliquants)
- Est-ce que les arêtes sont orientées ? (causalité)

Régulation \neq Co-expression

Gene Co-expression



Gene Regulation



Réseau de co-expression:

- Arêtes non orientées
- Arêtes reliant tous les types de gènes

Réseau de régulation:

- Arêtes orientées
- Arêtes partant des gènes régulateurs vers les gènes cibles

Interprétation d'un réseau de régulation

- Les réseaux de régulation sont **plus contraints** que les réseaux de co-expression, et portent une signification biologique **plus précise**
- Les réseaux de régulation recherchent plus de **causalité** dans l'explication des dépendances transcriptionnelles

Interprétation d'un réseau de régulation

- Les réseaux de régulation sont **plus contraints** que les réseaux de co-expression, et portent une signification biologique **plus précise**
- Les réseaux de régulation recherchent plus de **causalité** dans l'explication des dépendances transcriptionnelles

Attention aux interprétations

La causalité reste toute fois difficile à atteindre



Données, modélisation, inférence

Quel type de données pour quel réseau ?

Quel type de données pour quel réseau ?

- Réseau de co-expression
 - Mesure des niveaux d'expression (RNA-Seq)
- Réseau de régulation
 - Mesure des niveaux d'expression (RNA-Seq)
mais aussi :

Quel type de données pour quel réseau ?

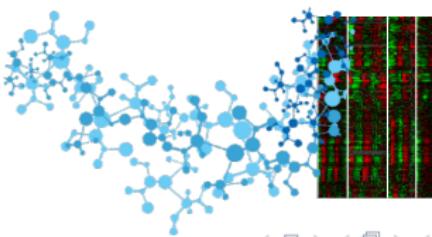
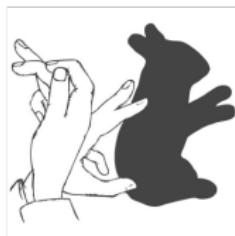
- Réseau de co-expression
 - Mesure des niveaux d'expression (RNA-Seq)
- Réseau de régulation
 - Mesure des niveaux d'expression (RNA-Seq)
mais aussi :
 - Recherche de motifs dans les séquences (génome + PWM)
 - Fixation de protéine (ChIP-Seq)
 - Accessibilité de la chromatine (ATAC-Seq)

2ème étape : l'estimation des paramètres

- Si les données d'expression à disposition contiennent $n_{g\acute{e}nes}$ gènes et n_{obs} conditions expérimentales:
 - les données à disposition sont de l'ordre de $n_{g\acute{e}nes} \times n_{obs}$
 - les données que nous aimerais obtenir, c.a.d les liens entre les gènes, sont de l'ordre de $n_{g\acute{e}nes} \times n_{g\acute{e}nes}$

2ème étape : l'estimation des paramètres

- **Si les données d'expression à disposition contiennent $n_{g\acute{e}nes}$ gènes et n_{obs} conditions expérimentales:**
 - les données à disposition sont de l'ordre de $n_{g\acute{e}nes} \times n_{obs}$
 - les données que nous aimeraient obtenir, c.a.d les liens entre les gènes, sont de l'ordre de $n_{g\acute{e}nes} \times n_{g\acute{e}nes}$
- **Le problème de la grande dimension ($n_{g\acute{e}nes} > n_{obs}$)**
Si $n_{g\acute{e}nes} > n_{obs}$, le volume d'information que nous essayons de reconstruire est plus grand que le volume d'information que nous avons à disposition! Et c'est très souvent le cas.



Méthodes classiques

Méthodes classiques pour l'inférence de GRN

Les **méthodes d'extraction de connaissances** (Domaine informatique) pour reconstruire un réseau à partir de :

base de connaissances

Méthodes classiques pour l'inférence de GRN

Les **méthodes d'extraction de connaissances** (Domaine informatique) pour reconstruire un réseau à partir de :

base de connaissances + algorithme de recherche

Méthodes classiques pour l'inférence de GRN

Les **méthodes d'extraction de connaissances** (Domaine informatique) pour reconstruire un réseau à partir de :

base de connaissances + algorithme de recherche

Connaissances :

- Données bibliographiques (publications scientifiques)
- Orthologie (base de données)
- ...

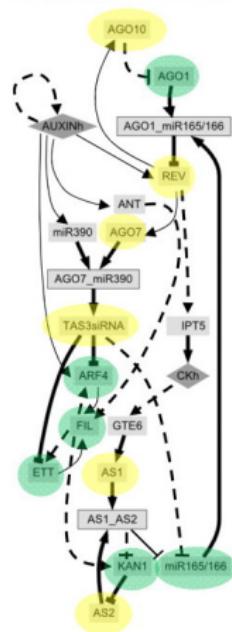
⇒ Reconstruction d'un réseau qui **rassemble des connaissances biologiques** issues de sources différentes.



Méthodes classiques pour l'inférence de GRN

Les méthodes d'extraction de connaissances (Informatique)

- **Réseau :**
 - Noeuds = gènes
 - Arêtes = relation observée dans la littérature scientifique, des bases de données, ...
 - **Avantages :**
 - Utilise les résultats établis par des experts et/ou issus de bases de données.
 - Rassemble et synthétise les résultats issus de différentes études.
 - **Limites** : long, se limite aux interactions déjà identifiées.



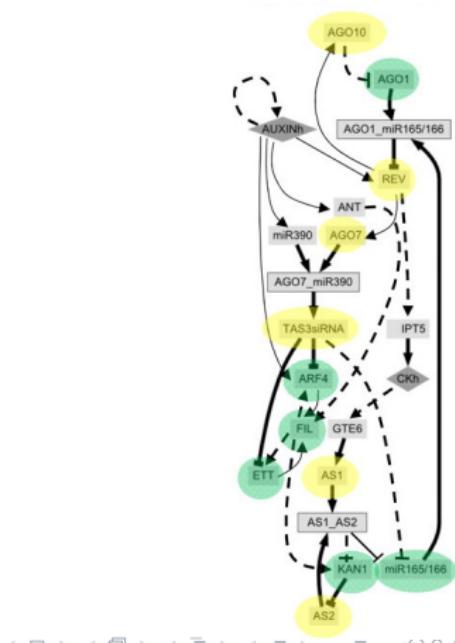


Méthodes classiques pour l'inférence de GRN

Les méthodes d'extraction de connaissances (Informatique)

- Méthodes les plus courantes

- traitement automatique des langues,
 - fouille de textes,
 - ontologie,
 - ...



Méthodes classiques pour l'inférence de GRN

Les **méthodes statistiques** pour reconstruire un réseau à partir de :

Méthodes classiques pour l'inférence de GRN

Les **méthodes statistiques** pour reconstruire un réseau à partir de :
données expérimentales

Méthodes classiques pour l'inférence de GRN

Les **méthodes statistiques** pour reconstruire un réseau à partir de :
données expérimentales + modèle probabiliste

Méthodes classiques pour l'inférence de GRN

Les **méthodes statistiques** pour reconstruire un réseau à partir de :

données expérimentales + modèle probabiliste + méthode d'estimation

Méthodes classiques pour l'inférence de GRN

Les **méthodes statistiques** pour reconstruire un réseau à partir de :

données expérimentales + modèle probabiliste + méthode d'estimation

- **Réseau :**
 - Noeuds = gènes
 - Arêtes = dépendances statistiques
 - **Avantages :** Permet de considérer des interactions avec des gènes encore inconnus. Permet de traiter les données recueillies dans des conditions spécifiques.
 - **Limites :** les arêtes sont des dépendances statistiques, il faudra les valider biologiquement.

Méthodes classiques pour l'inférence de GRN

Les méthodes statistiques

- Méthodes les plus courantes

- pour inférer un réseau de co-expression (non orienté)
 - Graphe de corrélation (Relevance Networks)
 - Graphe des corrélations partielles (GGM pour Gaussian Graphical Models)
 - Information mutuelle (Théorie de l'information)
 - pour inférer un réseau de régulation (orienté)
 - Réseaux Bayésiens (BN)
 - Modèles de régression (linéaire ou non)

Notations

On considère :

- $n_{gènes}$ **variables**
 - ⇒ On mesure simultanément le niveau d'expression de $n_{gènes}$ gènes
 - ⇒ On considère un vecteur aléatoire réel
 - $X = (X_1, \dots, X_i, \dots, X_{n_{gènes}})$ de dimension $n_{gènes}$, dont le i^{eme} élément correspond au niveau d'expression du gène i .
- n_{obs} **mesures**
 - ⇒ On observe n_{obs} fois le vecteur X (en général différentes conditions, différents points de temps, avec des réplicats pour chacun).
 - ⇒ On obtient un tableau de $n_{gènes}$ lignes et n_{obs} colonnes.



graphe des corrélations

1) Graphe des corrélations (Relevance network)

- Chaque gène est représenté par un noeud, qui représente son niveau d'expression : X_i .
- Une arête est tracée entre 2 gènes si et seulement si leurs expressions sont corrélées :

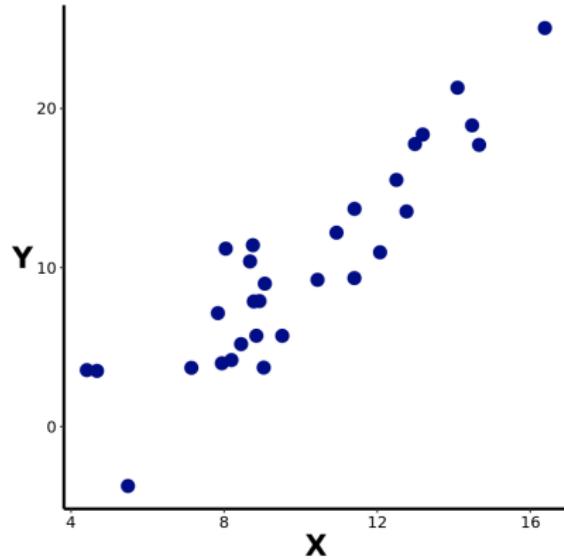
$$X_i - X_j \Leftrightarrow \text{cor}(X_i, X_j) \neq 0$$

	C_1	C_2	C_3	S_1	S_2	S_3	M_1	M_2	M_3	H_1	H_2	H_3
AT1G01050.1	846.8	840.2	837.2	836.1	818.8	792.2	802	761.4	719.9	922.2	969.3	930.8
AT1G01060.1	658.2	698.8	636.5	532.7	411	519.7	741.6	782.2	877.7	36.7	20.8	20.5



graphe des corrélations

Exemples : différentes valeurs de corrélation linéaire

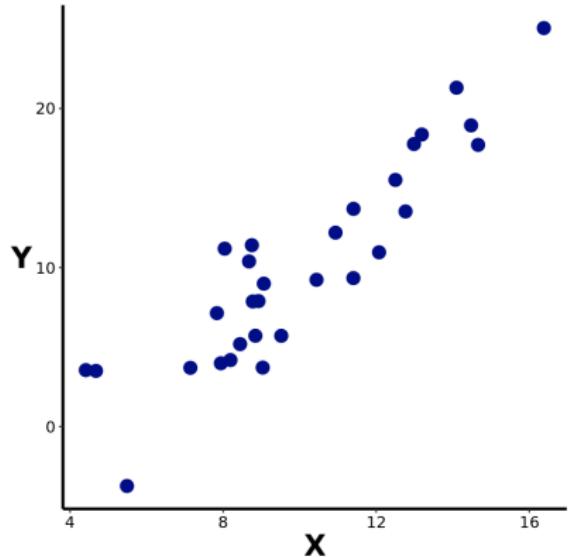


- $\text{cor}(X,Y) = 0.61 ?$



graphe des corrélations

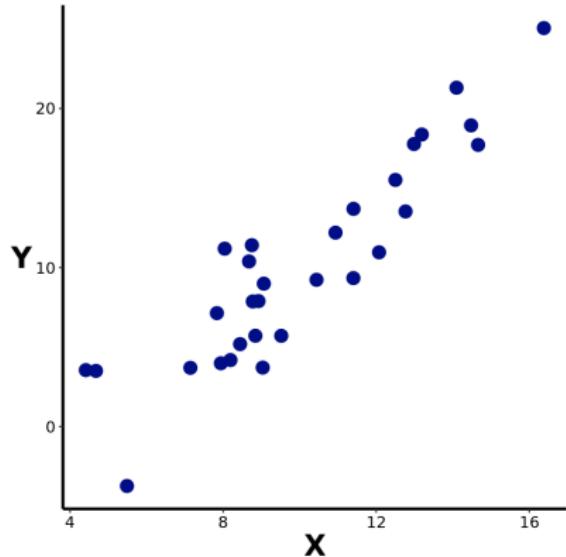
Exemples : différentes valeurs de corrélation linéaire



- $\text{cor}(X,Y) = 0.61 ?$
- $\text{cor}(X,Y) = 0.82 ?$

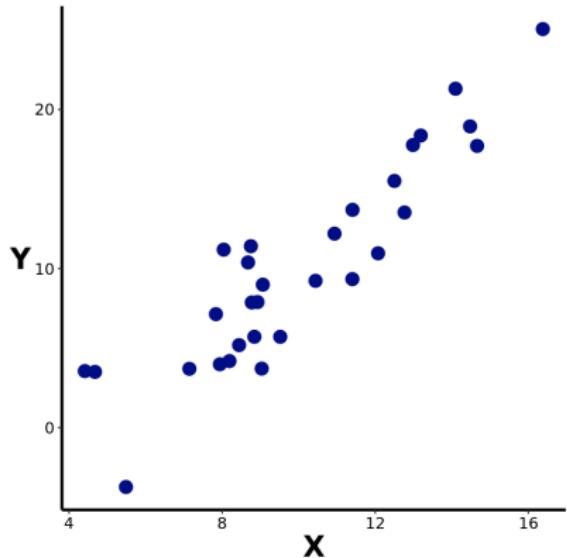
graphe des corrélations

Exemples : différentes valeurs de corrélation linéaire



- $\text{cor}(X,Y) = 0.61 ?$
- $\text{cor}(X,Y) = 0.82 ?$
- $\text{cor}(X,Y) = 0.91 ?$

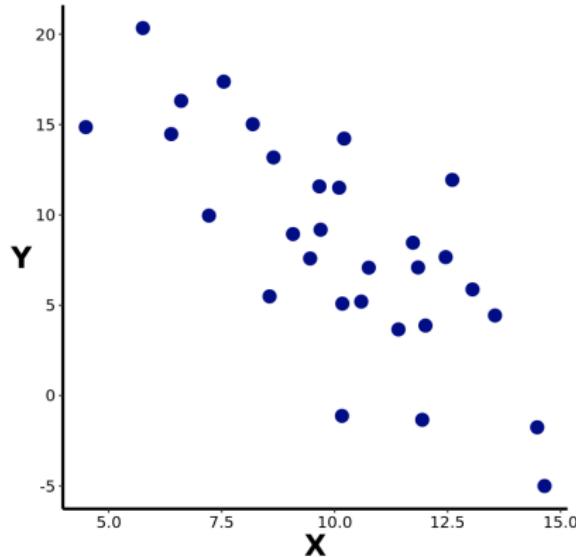
Exemples : différentes valeurs de corrélation linéaire



- $\text{cor}(X, Y) = 0.61 ?$
 - $\text{cor}(X, Y) = 0.82 ?$
 - $\text{cor}(X, Y) = 0.91 ?$
 - Réponse :
 $\text{cor}(X, Y) = 0.91$

graphe des corrélations

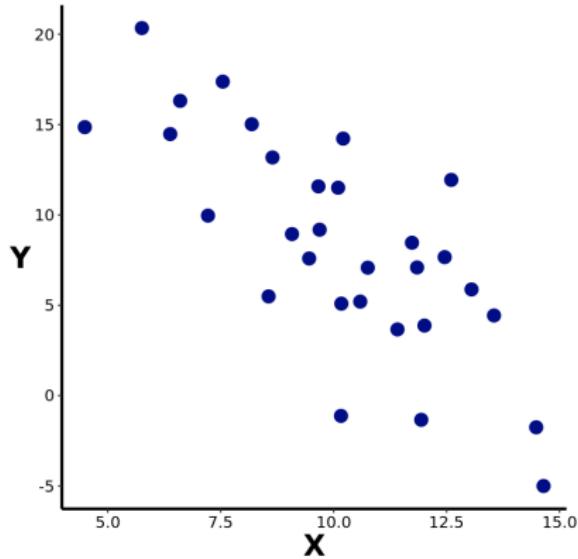
Exemples : différentes valeurs de corrélation linéaire



- $\text{cor}(X,Y) = -0.96 ?$

graphe des corrélations

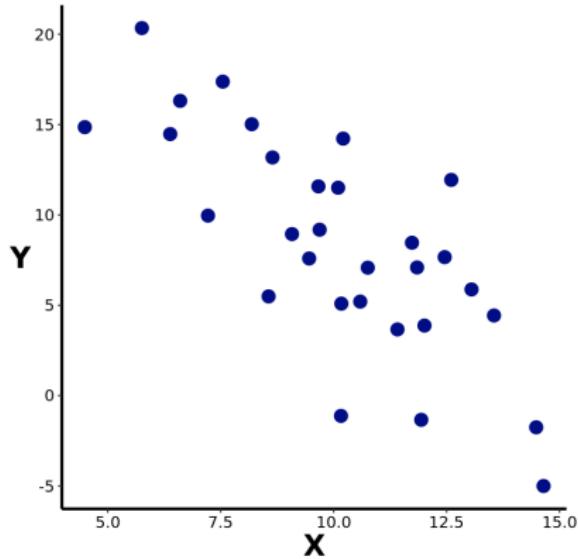
Exemples : différentes valeurs de corrélation linéaire



- $\text{cor}(X,Y) = -0.96 ?$
- $\text{cor}(X,Y) = -0.76 ?$

graphe des corrélations

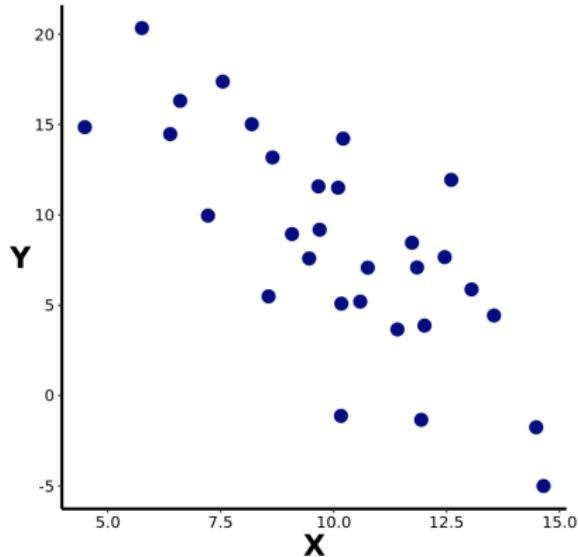
Exemples : différentes valeurs de corrélation linéaire



- $\text{cor}(X,Y) = -0.96 ?$
- $\text{cor}(X,Y) = -0.76 ?$
- $\text{cor}(X,Y) = -0.45 ?$

graphe des corrélations

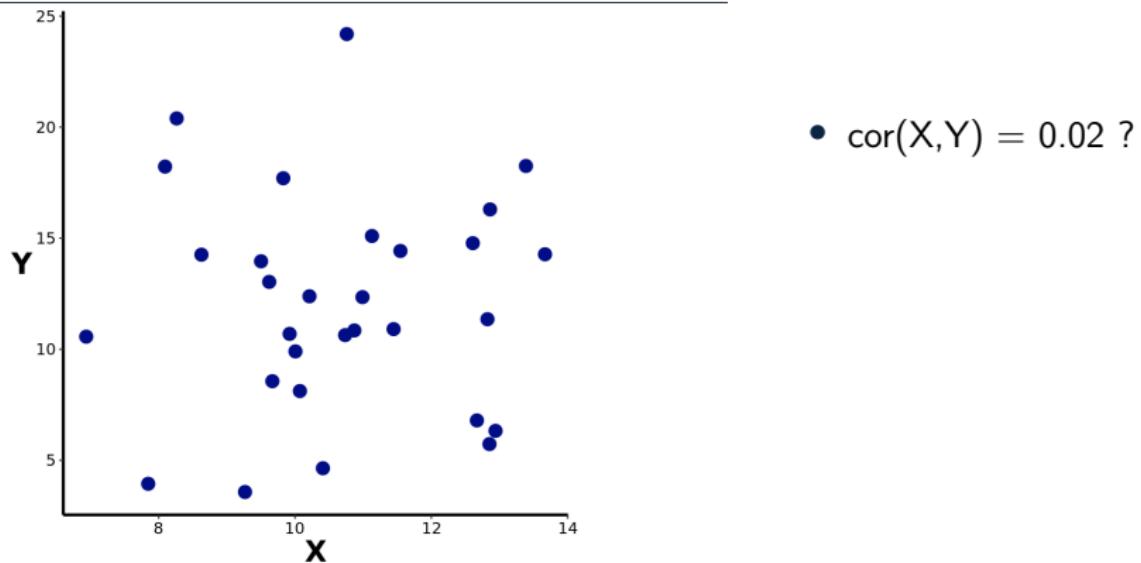
Exemples : différentes valeurs de corrélation linéaire



- $\text{cor}(X,Y) = -0.96 ?$
- $\text{cor}(X,Y) = -0.76 ?$
- $\text{cor}(X,Y) = -0.45 ?$
- **Réponse :**
 $\text{cor}(X,Y) = -0.76$

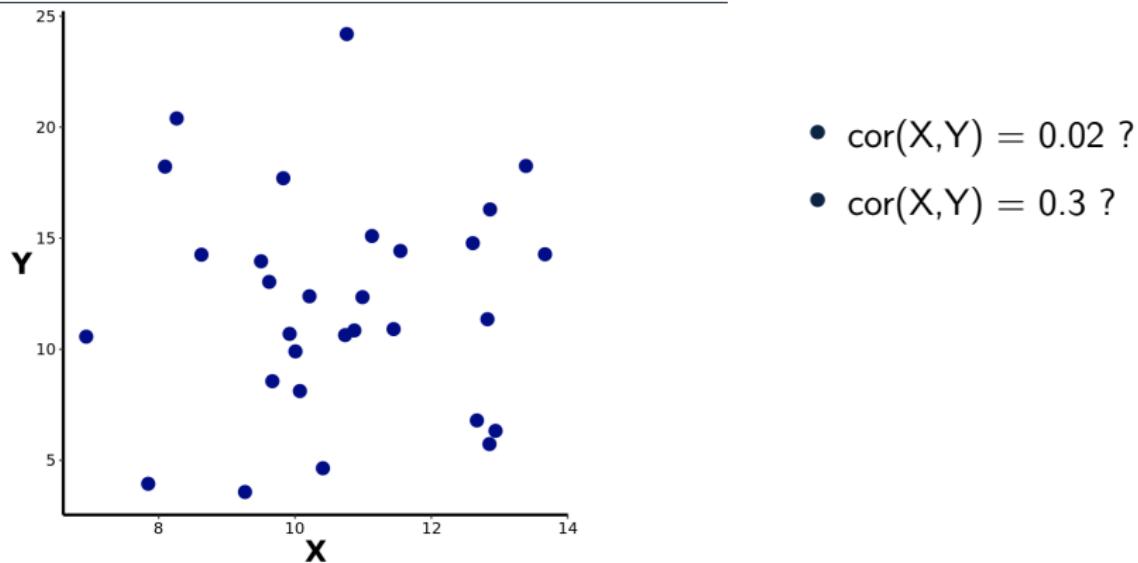
graphe des corrélations

Exemples : différentes valeurs de corrélation linéaire



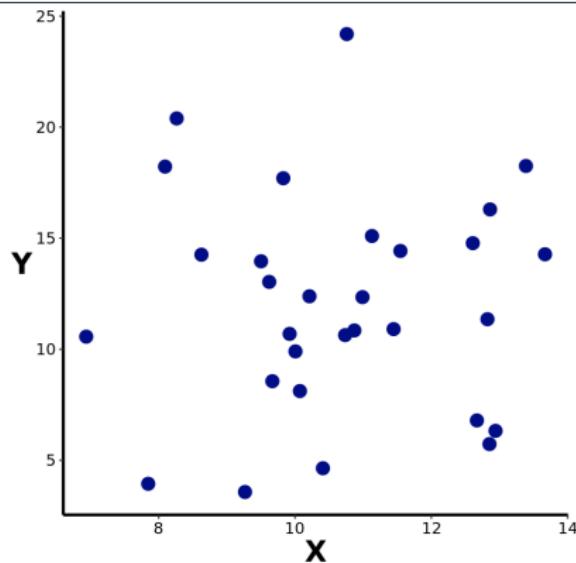
graphe des corrélations

Exemples : différentes valeurs de corrélation linéaire



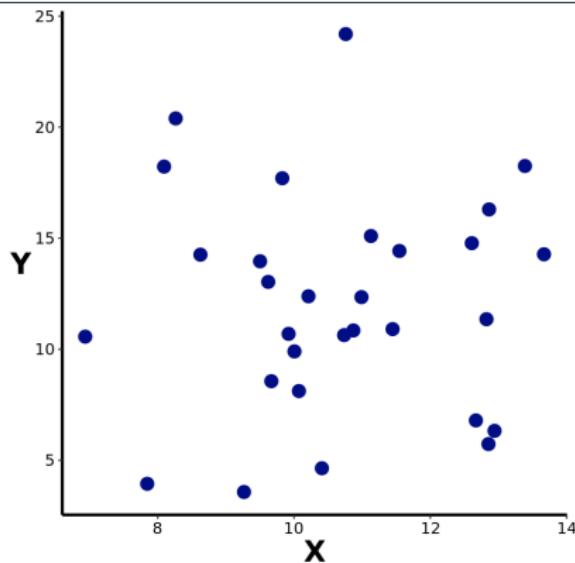
graphe des corrélations

Exemples : différentes valeurs de corrélation linéaire



- $\text{cor}(X,Y) = 0.02 ?$
- $\text{cor}(X,Y) = 0.3 ?$
- $\text{cor}(X,Y) = 0.6 ?$

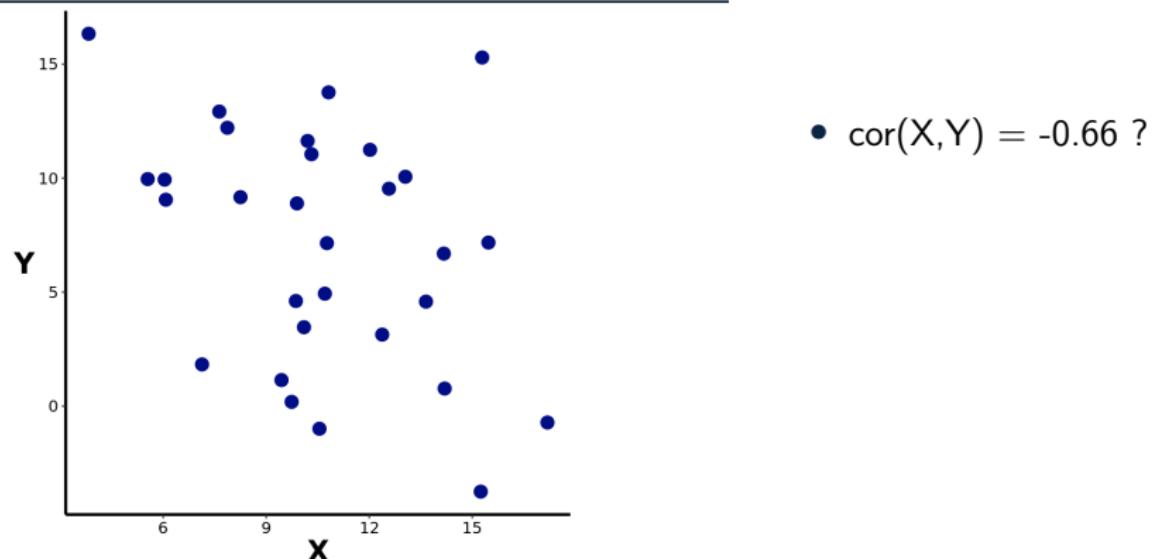
Exemples : différentes valeurs de corrélation linéaire



- $\text{cor}(X,Y) = 0.02 ?$
- $\text{cor}(X,Y) = 0.3 ?$
- $\text{cor}(X,Y) = 0.6 ?$
- **Réponse :**
 $\text{cor}(X,Y) = 0.02$

graphe des corrélations

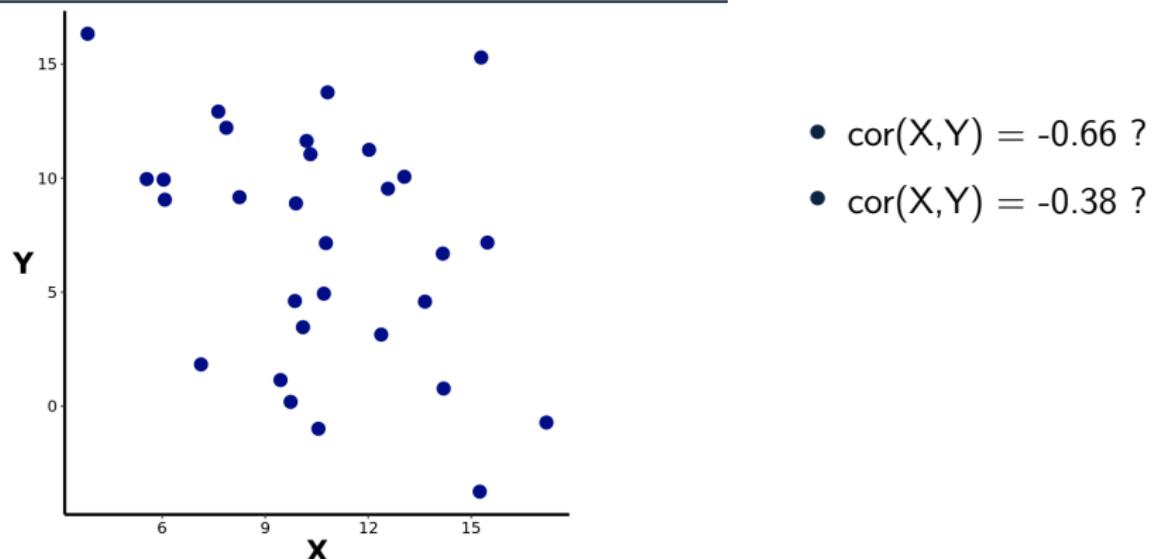
Exemples : différentes valeurs de corrélation linéaire



Pour d'autres simulations en ligne :

https://shiny.zoology.ubc.ca/whitlock/Guessing_correlation/

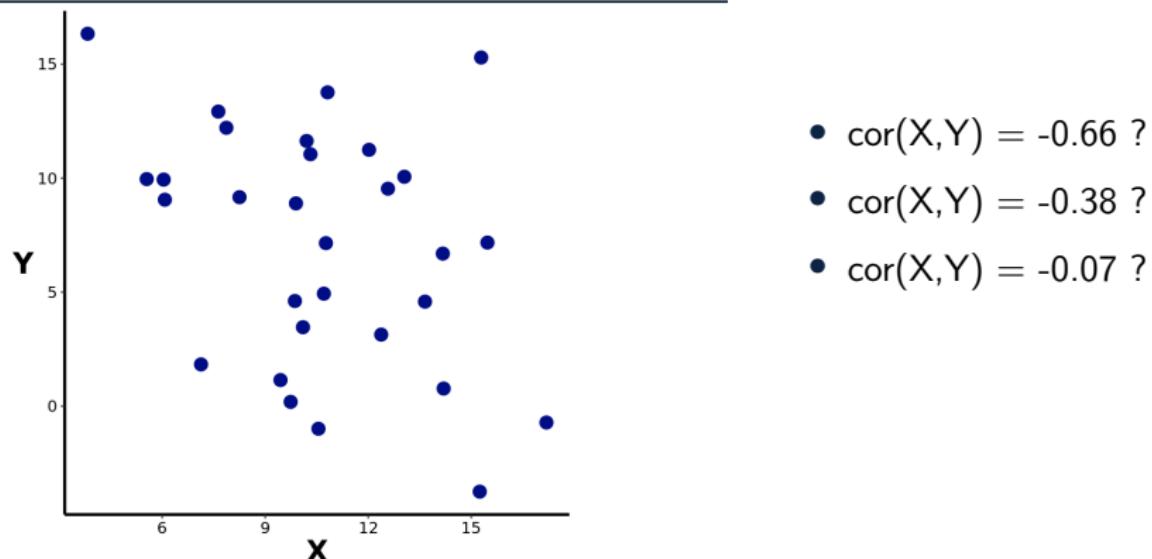
Exemples : différentes valeurs de corrélation linéaire



Pour d'autres simulations en ligne :

https://shiny.zoology.ubc.ca/whitlock/Guessing_correlation/

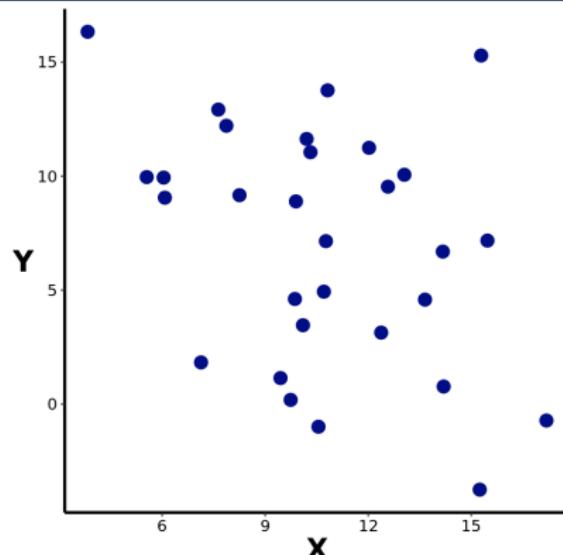
Exemples : différentes valeurs de corrélation linéaire



Pour d'autres simulations en ligne :

https://shiny.zoology.ubc.ca/whitlock/Guessing_correlation/

Exemples : différentes valeurs de corrélation linéaire



- $\text{cor}(X,Y) = -0.66 ?$
- $\text{cor}(X,Y) = -0.38 ?$
- $\text{cor}(X,Y) = -0.07 ?$
- Réponse :
 $\text{cor}(X,Y) = -0.38$

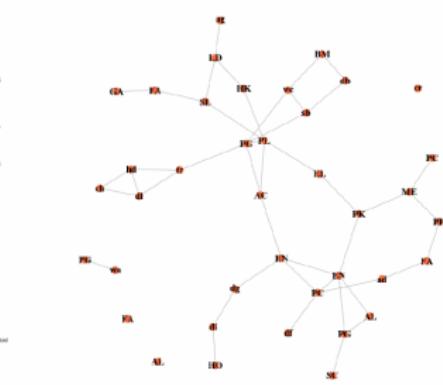
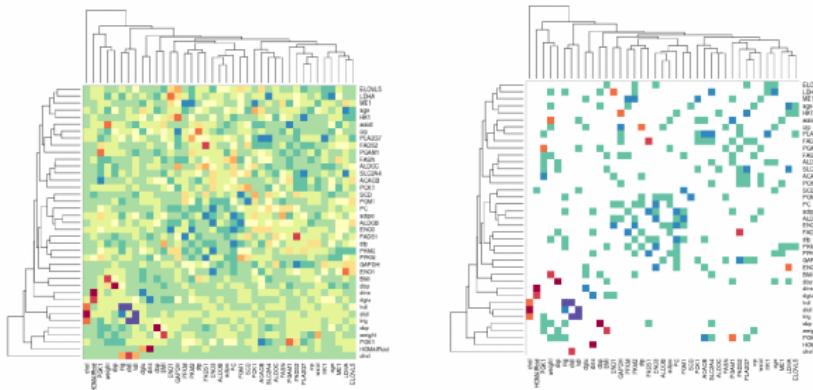
Pour d'autres simulations en ligne :

https://shiny.zoology.ubc.ca/whitlock/Guessing_correlation/

1) Graphe des corrélations (Relevance network)

En pratique :

- ① Calcul de la matrice de corrélations de taille $n_{gènes} \times n_{gènes}$ (heatmap)
- ② Choix d'un seuil limite –> sélection des arêtes
- ③ Construction du réseau

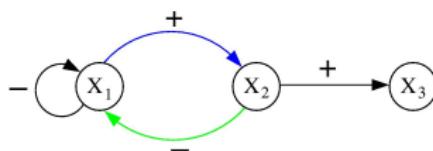


1) Graphe des corrélations (Relevance network)

Limites :

- Comment fixer le seuil de présence ou absence d'arête ?
- Identifie également les corrélations 'indirectes'

Exemple motif de régulation :



Quel graphe des corrélations serait attendu pour ce motif ?

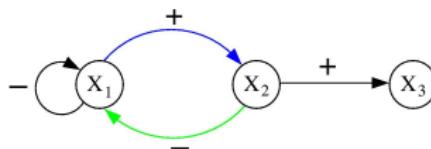
graphe des corrélations

1) Graphe des corrélations (Relevance network)

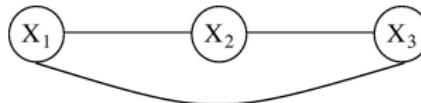
Limites :

- Comment fixer le seuil de présence ou absence d'arête ?
- Identifie également les corrélations 'indirectes'

Exemple motif de régulation :



Quel graphe des corrélations serait attendu pour ce motif ?





graphe des corrélations

1) Graphe des corrélations (Relevance network)

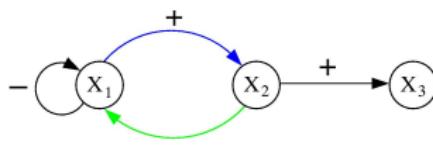
Conclusions :

- Représente la co-expression
- Permet d'identifier de groupes de gènes impliqués dans un même processus biologique

... mais pas de sélectionner uniquement les interactions 'directes'.

2) Graphe des corrélations partielles

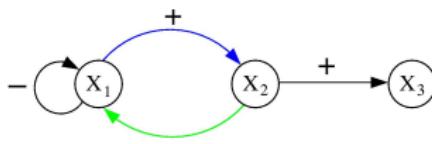
- Pour savoir si, derrière une liaison entre X et Y se cache en fait :
 - une liaison entre X et Z ,
 - et une autre entre Y et Z ,on calcule la **corrélation partielle** $\text{cor}(X, Y|Z)$
- Exemple motif de régulation :



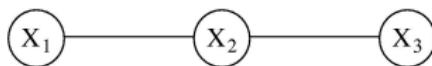
- Quel graphe des corrélations partielles serait attendu pour ce motif ?

2) Graphe des corrélations partielles

- Pour savoir si, derrière une liaison entre X et Y se cache en fait :
 - une liaison entre X et Z ,
 - et une autre entre Y et Z ,
 on calcule la **corrélation partielle** $\text{cor}(X, Y|Z)$
- Exemple motif de régulation :



- Quel graphe des corrélations partielles serait attendu pour ce motif ?



2) Graphe des corrélations partielles

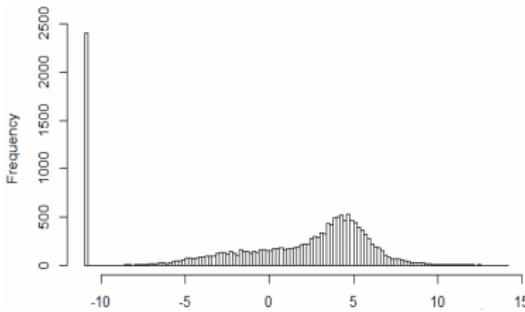
- Repose sur une **hypothèse gaussienne (ou normale)**:

$$X = (X_1, \dots, X_{n_{génomes}}) \sim \mathcal{N}(\mu, \Sigma)$$

car une corrélation nulle entre 2 variables suivant une loi normale implique que ces 2 variables sont indépendantes.

(\rightsquigarrow Modèles Graphiques Gaussien (GGM))

- En pratique : **Transformation log** des données de comptage (RNA-Seq)



2) Graphe des corrélations partielles

- Graphe des corrélations partielles (Définition) :

Présence de l'arête $X_i - X_j \Leftrightarrow \text{cor}(X_i, X_j | (X_k)_{k \neq i,j}) \neq 0$

- On calcule les corrélations partielles grâce à la matrice de concentration $\Theta = \Sigma^{-1}$, qui est l'inverse de la matrice de variance-covariance Σ (de taille $n_{gènes} \times n_{gènes}$) , et qui satisfait :

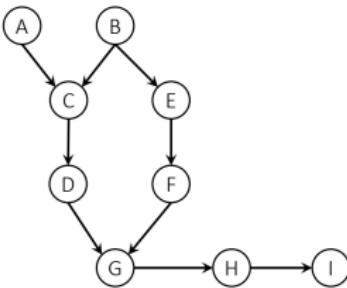
$$\text{cor}(X_i, X_j | (X_k)_{k \neq i,j}) = -\frac{\theta_{ij}}{\sqrt{\theta_{ii}\theta_{jj}}}$$

N.B. : Plus les mesures sont de qualité et en nombre (n_{obs}) élevé, mieux on estime Σ et Σ^{-1} .

3) Réseaux bayésiens

- Comme les GGM, les réseaux bayésiens sont des modèles probabilistes qui visent à représenter des dépendances conditionnelles.
- Un RB est défini par un graphe orienté **acyclique** (ou DAG pour Directed Acyclic Graph)

Exemple :



- Interprétation: un RB permet de déduire des relations d'indépendances conditionnelles entre les variables considérées.



graphe des corrélations

3) Réseaux bayésiens

- **Limites :**

- Plus difficile à modéliser (pas de cycle)
- Plus difficile à estimer (2 orientations possibles pour chaque arête + contrainte acyclique)
- **Attention à l'interprétation** de l'orientation des arêtes.



graphe des corrélations

3) Réseaux bayésiens

- **Limites :**

- Plus difficile à modéliser (pas de cycle)
- Plus difficile à estimer (2 orientations possibles pour chaque arête + contrainte acyclique)
- **Attention à l'interprétation** de l'orientation des arêtes.
 - ~~ Des RBs différents peuvent représenter les mêmes indépendances conditionnelles.

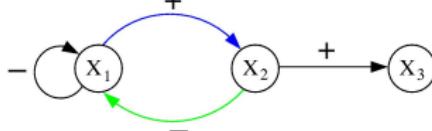
3) Réseaux bayésiens

- **Limites :**

- Plus difficile à modéliser (pas de cycle)
- Plus difficile à estimer (2 orientations possibles pour chaque arête + contrainte acyclique)
- **Attention à l'interprétation** de l'orientation des arêtes.

~~ Des RBs différents peuvent représenter les mêmes indépendances conditionnelles.

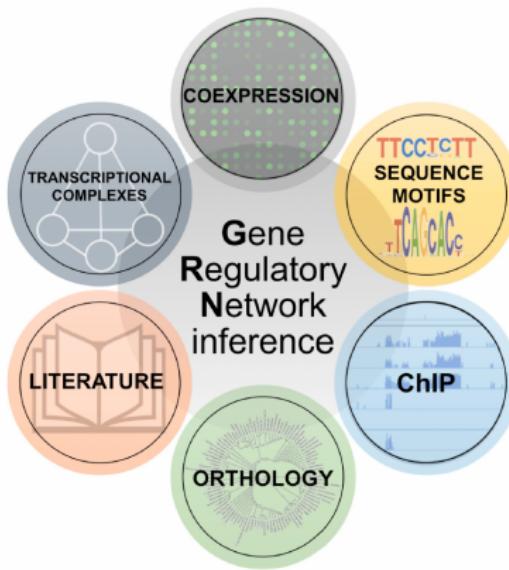
Ex : Le motif de régulation ci-dessous peut être représenté par plusieurs DAG équivalents en termes d'indépendances conditionnelles.



- Assez peu utilisé à l'échelle génome wide.

Pour aller plus loin sur l'inférence de GRN

Un article de review : Gene regulatory network inference resources: A practical overview. Mercatellia et al. (2020).



Utiliser un *a priori* biologique

- pour **réduire la dimension** (seuls certains gènes sont considérés)
 - gènes connus pour être impliqués dans le processus étudié
 - gènes différentiellement exprimés dans la ou les condition(s) étudiée(s)
 - gènes situés dans une région de l'accessibilité de la chromatine
- pour **orienter les arêtes** (causalité)
 - selon le temps : $X_t \rightarrow X_{t+1}$
 - selon le sens : Régulateur connu \rightarrow Gène cible
(chez la plante *A. thaliana* environ 2000 régulateurs vs 30 000 gènes)
 - selon la présence de motif (PWM) dans la région promotrice du gène cible : Régulateur associé au motif \rightarrow Gène cible

N.B. : et on peut combiner les deux !

Se ramener à un problème de régression

Orientier les arêtes *a priori* permet de se ramener à un problème de **régression**.

Se ramener à un problème de régression

Orientier les arêtes *a priori* permet de se ramener à un problème de **régression**.

Il s'agit d'expliquer, **pour chaque gène cible**, son niveau d'expression à partir du sous-ensemble de régulateurs potentiels (déterminé par l'*a priori*)

:

Se ramener à un problème de régression

Orientier les arêtes *a priori* permet de se ramener à un problème de **régression**.

Il s'agit d'expliquer, **pour chaque gène cible**, son niveau d'expression à partir du sous-ensemble de régulateurs potentiels (déterminé par l'*a priori*)

:

- les gènes observés au temps précédent

Se ramener à un problème de régression

Orientier les arêtes *a priori* permet de se ramener à un problème de **régression**.

Il s'agit d'expliquer, **pour chaque gène cible**, son niveau d'expression à partir du sous-ensemble de régulateurs potentiels (déterminé par l'*a priori*)

:

- les gènes observés au temps précédent
- les gènes connus comme facteurs de transcription

Se ramener à un problème de régression

Orientier les arêtes *a priori* permet de se ramener à un problème de **régression**.

Il s'agit d'expliquer, **pour chaque gène cible**, son niveau d'expression à partir du sous-ensemble de régulateurs potentiels (déterminé par l'*a priori*)

:

- les gènes observés au temps précédent
- les gènes connus comme facteurs de transcription
- les gènes dont le motif de fixation est situé dans la région promotrice du gène cible.

Cela revient à estimer $n_{g\acute{e}nes}$ modèles de régression.



Inférence de GRN par régression

Qu'est-ce qu'une régression?

La régression est une procédure statistique visant à établir la relation entre

- ➊ Une **variable d'intérêt**, également désignée comme variable réponse ou variable dépendante
- ➋ Une ou plusieurs **variables prédictives**, également désignées comme covariables, prédicteurs, ou variables indépendantes



Qu'est-ce qu'une régression?

La régression est une procédure statistique visant à établir la relation entre

- ① Une **variable d'intérêt** Y : vecteur d'observations de la réponse
- ② Une ou plusieurs **variables prédictives** X : vecteur ou matrice de prédicteurs pour chaque observation

Le lien entre X et Y se fait au moyen d'une fonction f , estimée lors de la procédure de régression, telle que:

$$Y_i = \mathbf{f}(X_i) + \epsilon_i$$

Avec i allant de 1 à N , et ϵ_i l'erreur (ou résidus) du modèle, soit la variation de Y non expliquée par les prédicteurs X .



Exemple de régression : la biomasse

La régression est une procédure statistique visant à établir la relation entre

- ➊ Une **variable d'intérêt** Y : la **biomasse** d'Arabidopsis thaliana



- ➋ Une ou plusieurs **variables prédictives** X : **L'apport en nitrate, la température, l'humidité, etc.**

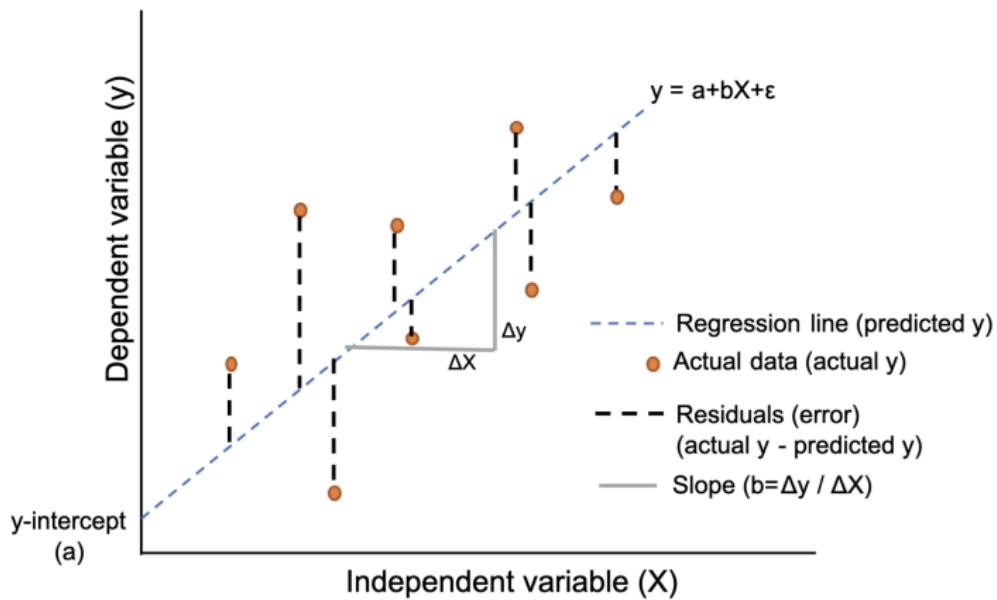
$$\text{Biomasse}_i = \mathbf{f}(\text{Nitrate}_i) + \epsilon_i$$

Avec i allant de 1 à N correspondant à chacune des plantes du jeu de données, et ϵ_i la variation de biomasse non expliquée par l'apport en nitrate, la température, l'humidité, etc.



Représentation graphique

Cas d'une régression **linéaire** à un prédicteur : $Y_i = a + bX_i + \epsilon_i$



Régression simple et régression multiple

Cas d'une régression **linéaire** à deux prédicteurs (sans interaction):

$$\text{Biomasse}_i = \mathbf{a} + \mathbf{b} \text{Nitrate}_i + \mathbf{c} \text{Temperature}_i + \epsilon_i$$



Régression simple et régression multiple

Cas d'une régression **linéaire** à deux prédicteurs (sans interaction):

$$\text{Biomasse}_i = \mathbf{a} + \mathbf{b} \text{Nitrate}_i + \mathbf{c} \text{Temperature}_i + \epsilon_i$$

Cas d'une régression **linéaire** à deux prédicteurs (avec interaction):

$$\text{Biomasse}_i = \mathbf{a} + \mathbf{b} \text{Nitrate}_i + \mathbf{c} \text{Temperature}_i + \mathbf{d} \text{Temperature}_i * \text{Nitrate}_i + \epsilon_i$$

Remarques

Avantages de la régression :

- Possibilité de prédire la variable d'intérêt pour un nouveau set de variables à l'outcome inconnu
- Quantification du pouvoir prédicteur de chaque variable d'entrée (ex: valeur du coefficient, et test sur la significativité de ce coefficient.)



Remarques

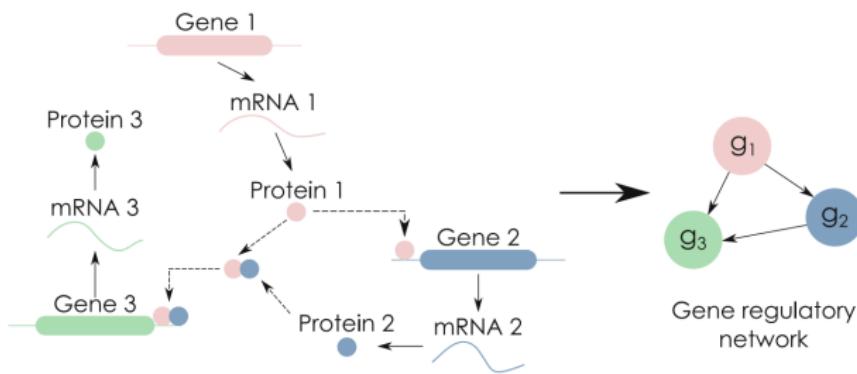
Avantages de la régression :

- Possibilité de prédire la variable d'intérêt pour un nouveau set de variables à l'outcome inconnu
- Quantification du pouvoir prédicteur de chaque variable d'entrée (ex: valeur du coefficient, et test sur la significativité de ce coefficient.)

Extensions de la régression linéaire utiles à la suite du cours :

- On peut ajouter des termes de pénalisation lors de l'estimation des coefficients afin de faire de la sélection de variables en grande dimension (lasso)
- La fonction f n'est pas nécessairement une combinaison linéaire du type $a + bX_1 + cX_2$, mais peut prendre la forme des arbres de régression, des fonctions non linéaires plus adaptées à des questions biologiques complexes

Principe de l'application aux réseaux de gènes



g_{1i}, g_{2i} : niveaux d'expression des régulateurs transcriptionnels dans la condition i
 g_{3i} : niveaux d'expression d'un gène cible dans la condition i

$$g_{3i} = \mathbf{f}(g_{1i}, g_{2i}) + \epsilon_i$$



Principe de l'application aux réseaux de régulation

regulators_i : niveaux d'expression des régulateurs transcriptionnels dans la condition i

target_i : niveaux d'expression d'un gène cible dans la condition i

$$\text{target}_i = \mathbf{f}(\text{regulators}_i) + \epsilon_i$$

La procédure de construction de réseau est la suivante :

- ① Pour chaque gène du jeu de données, ajuster à partir des valeurs d'expression la fonction \mathbf{f}
- ② Extraire de \mathbf{f} les scores (ou valeur d'influence, importance, pouvoir prédictif) des régulateurs sur chaque gène du jeu de données
- ③ Sélectionner les scores régulateurs-gènes cibles les plus forts pour construire le réseau final



Principe de l'application aux réseaux de régulation

$regulators_i$: niveaux d'expression des régulateurs transcriptionnels dans la condition i
 $target_i$: niveaux d'expression d'un gène cible dans la condition i

$$target_i = f(regulators_i) + \epsilon_i$$

Les méthodes basées sur la régression pour inférerer des GRN se différencient par leur choix de modélisation pour f . Deux sont présentées dans ce cours :

- **TIGRESS** : Régression linéaire pénalisée avec sélection de stabilité [Haury et al., 2012]
- **GENIE3** : Arbres de régression en random forests [Huynh-Thu et al., 2010]

2 exemples de méthodes



Exemple de méthodes : TIGRESS et GENIE3

TIGRESS : approche basée sur la régression linéaire

TIGRESS fait le choix de modélisation suivant : l'expression d'un gène cible peut être modélisée par une **combinaison linéaire** de l'expression des facteurs de transcription :



Exemple de méthodes : TIGRESS et GENIE3

TIGRESS : approche basée sur la régression linéaire

TIGRESS fait le choix de modélisation suivant : l'expression d'un gène cible peut être modélisée par une **combinaison linéaire** de l'expression des facteurs de transcription :

$$\text{target}_i = \beta_{\text{target},1} \cdot \text{TF1}_i + \beta_{\text{target},2} \cdot \text{TF2}_i + \dots + \beta_{\text{target},M} \cdot \text{TFM}_i + \epsilon_i$$

Avec TFM_i le niveau d'expression du TF numéro M dans la condition i .



Exemple de méthodes : TIGRESS et GENIE3

TIGRESS : approche basée sur la régression linéaire

TIGRESS fait le choix de modélisation suivant : l'expression d'un gène cible peut être modélisée par une **combinaison linéaire** de l'expression des facteurs de transcription :

$$\text{target}_i = \beta_{\text{target},1} \cdot \text{TF1}_i + \beta_{\text{target},2} \cdot \text{TF2}_i + \dots + \beta_{\text{target},M} \cdot \text{TFM}_i + \epsilon_i$$

Avec TFM_i le niveau d'expression du TF numéro M dans la condition i .

Problème : il faut refléter la sparsité du problème biologique

L'expression d'un gène est sensée être expliquée par un nombre limité de TFs, et non tous les TFs du jeu de données → **LARS** (Least-angle regression)



Exemple de méthodes : TIGRESS et GENIE3

TIGRESS : Modéliser la sparsité

Problème : il faut refléter la sparsité du problème biologique

L'expression d'un gène est sensée être expliquée par un nombre limité de TFs, et non tous les TFs du jeu de données → **LARS** (Least-angle regression)

La méthode LARS fonctionne (dans le principe) comme suit :

- ① Commencer par le modèle nul : $target_i = \alpha + \epsilon_i$
- ② Choisir le TF_j pour lequel la corrélation à $target$ est maximale :
 $target_i = \alpha + \beta_{target,j} \cdot TF_j + \epsilon_i$
- ③ Continuer à ajouter des TFs en choisissant à chaque fois le TF maximisant la prédiction de $target$
- ④ S'arrêter lorsque l'on a atteint le nombre de TFs prédicteurs jugé suffisant, ici $L < M$.

On termine alors avec le modèle suivant, constitué de uniquement L TFs contre M , sans contrainte de sparsité : $target_i = \beta_{target,1} \cdot TF_1 + \beta_{target,2} \cdot TF_2 + \dots + \beta_{target,L} \cdot TF_L + \epsilon_i$



Exemple de méthodes : TIGRESS et GENIE3

TIGRESS : étapes de la procédure

	Response	Predictors		
	Target gene	TF 1	...	TF N
Experimental conditions				
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				
11				
12				
13				
14				
15				
16				
17				
18				
19				
20				
21				
22				
23				
24				
25				
26				
27				
28				
29				
30				
31				
32				
33				
34				
35				
36				
37				
38				
39				
40				
41				
42				
43				
44				
45				
46				
47				
48				
49				
50				
51				
52				
53				
54				
55				
56				
57				
58				
59				
60				
61				
62				
63				
64				
65				
66				
67				
68				
69				
70				
71				
72				
73				
74				
75				
76				
77				
78				
79				
80				
81				
82				
83				
84				
85				
86				
87				
88				
89				
90				
91				
92				
93				
94				
95				
96				
97				
98				
99				
100				

$$X_g = f_g(X_{\mathcal{F}_g}) = \sum_{t \in \mathcal{F}_g} \beta_{t,g} X_t$$



Exemple de méthodes : TIGRESS et GENIE3

TIGRESS : étapes de la procédure

	Response	Predictors		
	Target gene	TF 1	...	TF N
Experimental conditions				
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				

LARS : add **L predictors** iteratively,
each time choosing which one
helps to explains the most X_{target}

$$X_g = f_g \left(X_{\mathcal{F}_g} \right) = \sum_{i \in \mathcal{F}_g} \beta_i X_i$$

- Top L predictors: $TF_2 \dots TF_8 \dots TF_5$
- N-L TFs not selected



Exemple de méthodes : TIGRESS et GENIE3

TIGRESS : étapes de la procédure

	Response	Predictors		
	Target gene	TF 1	...	TF N
Experimental conditions				

Stability selection : Runs LARS R times, each time slightly randomly perturbing the expression data

LARS : add **L predictors** iteratively, each time choosing which one helps to explains the most X_{target}

$$X_g = f_g(X_{\mathcal{T}_g}) = \sum_{t \in \mathcal{T}_g} \beta_{t,g} X_t$$

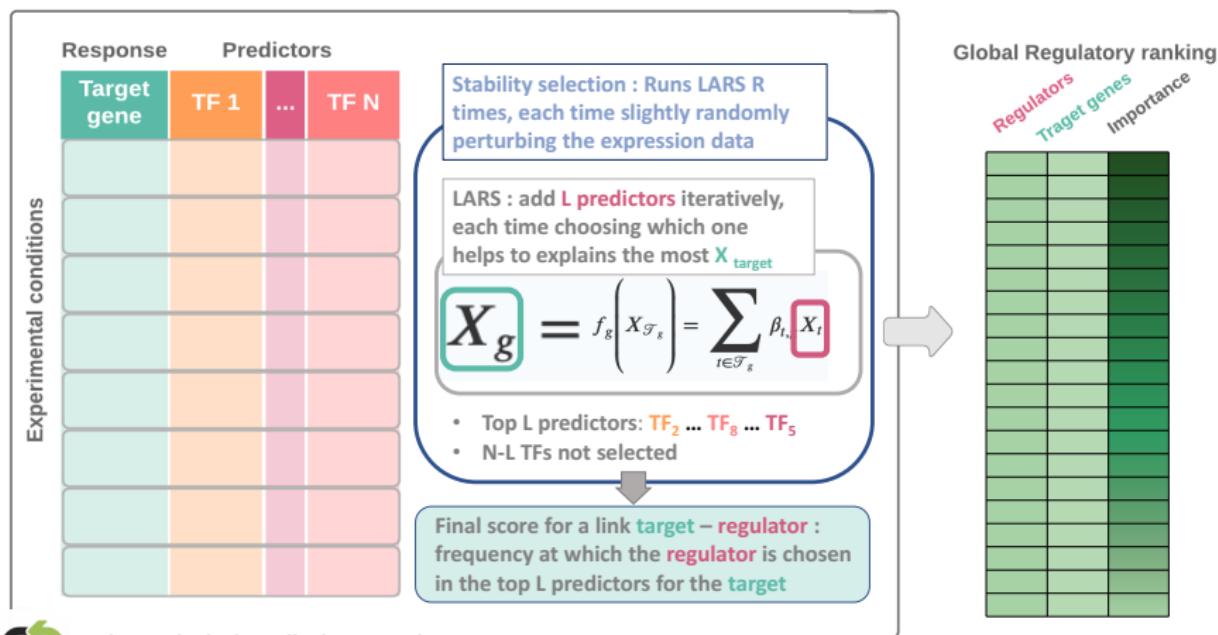
- Top L predictors: TF₂ ... TF₈ ... TF₅
- N-L TFs not selected

Final score for a link **target – regulator** : frequency at which the **regulator** is chosen in the top L predictors for the **target**



Exemple de méthodes : TIGRESS et GENIE3

TIGRESS : étapes de la procédure



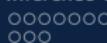
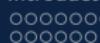
Each gene in the input list becomes the target



Exemple de méthodes : TIGRESS et GENIE3

GENIE3 : approche basée sur les arbres de régression

GENIE3 fait le choix de modélisation suivant : l'expression d'un gène cible peut être modélisée par une **combinaison non linéaire** de l'expression des facteurs de transcription :



Exemple de méthodes : TIGRESS et GENIE3

GENIE3 : approche basée sur les arbres de régression

GENIE3 fait le choix de modélisation suivant : l'expression d'un gène cible peut être modélisée par une **combinaison non linéaire** de l'expression des facteurs de transcription :

$$\text{target}_i = \text{RandomForest}(\text{TF}_i) + \epsilon_i$$

(On n'a pas de formulation mathématique pour le modèle d'un random forest, qui fonctionne très différemment d'une régression linéaire)

Avec TF_i le niveau d'expression de tous les TFs du jeu de données dans la condition i .

Avantages par rapport au modèle linéaire

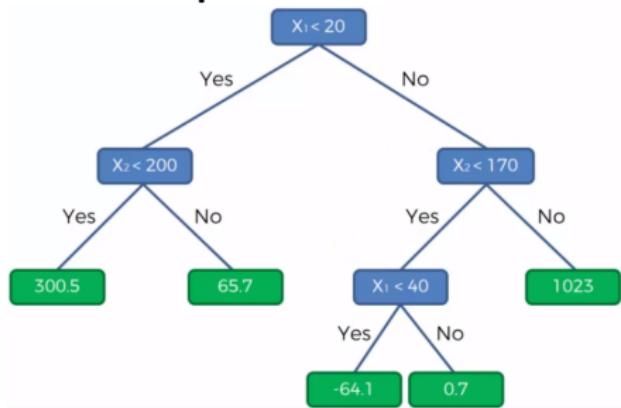
- Peut modéliser des non linéarités dans l'influence de l'expression des régulateurs (ex: le carré de l'expression d'un régulateur, etc)
- Peut modéliser des relations de coopération et d'interactions entre TFs



Exemple de méthodes : TIGRESS et GENIE3

GENIE3 : approche basée sur les arbres de régression

Un arbre de régression est construit en choisissant **des seuils et conditions sur les variables prédictives**.



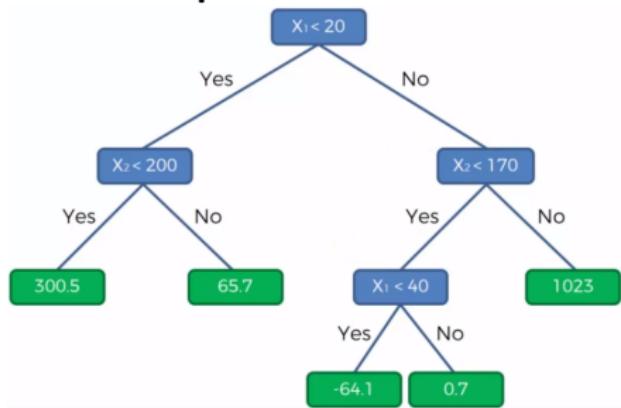
Ajustement d'un arbre de régression

- ① Choisir la variable et la condition sur cette variable qui permettent de discriminer au mieux les valeurs de la réponse (la variance de la réponse est diminuée)
- ② Réitérer en créant de nouvelles branches, jusqu'à épuisement des variables, ou atteinte de la profondeur d'arbre maximale

Exemple de méthodes : TIGRESS et GENIE3

GENIE3 : approche basée sur les arbres de régression

Un arbre de régression est construit en choisissant **des seuils et conditions sur les variables prédictives**.



Ajustement d'un arbre de régression

- ① Choisir la variable et la condition sur cette variable qui permettent de discriminer au mieux les valeurs de la réponse (la variance de la réponse est diminuée)
- ② Réitérer en créant de nouvelles branches, jusqu'à épuisement des variables, ou atteinte de la profondeur d'arbre maximale

Random Forest : un grand nombre d'arbres de régression sont ajustés sur des données échantillonées légèrement différemment les unes des autres → leur consensus permet plus de robustesse dans les prédictions (apprentissage ensembliste)



Exemple de méthodes : TIGRESS et GENIE3

GENIE3: étapes de la procédure

Ranking the regulators according to their **relevance for predicting** the other genes expression

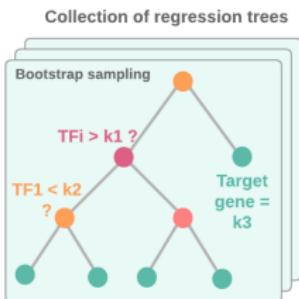
Response	Predictors		
Target gene	TF 1	...	TF N
Experimental conditions			



Exemple de méthodes : TIGRESS et GENIE3

GENIE3: étapes de la procédure

Ranking the regulators according to their **relevance for predicting** the other genes expression

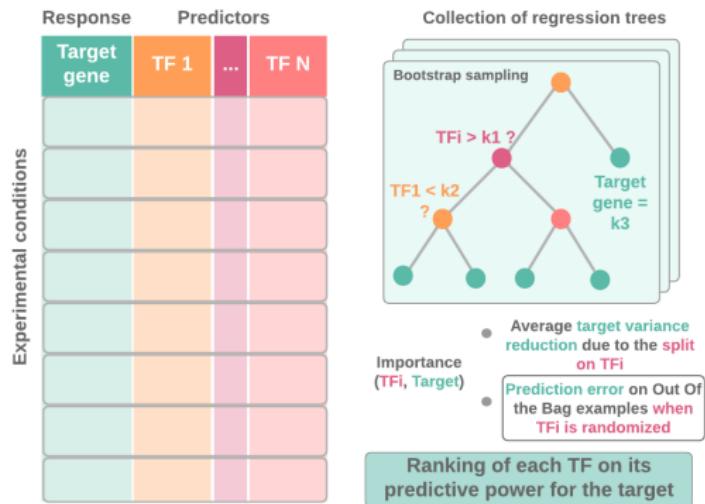




Exemple de méthodes : TIGRESS et GENIE3

GENIE3: étapes de la procédure

Ranking the regulators according to their relevance for predicting the other genes expression

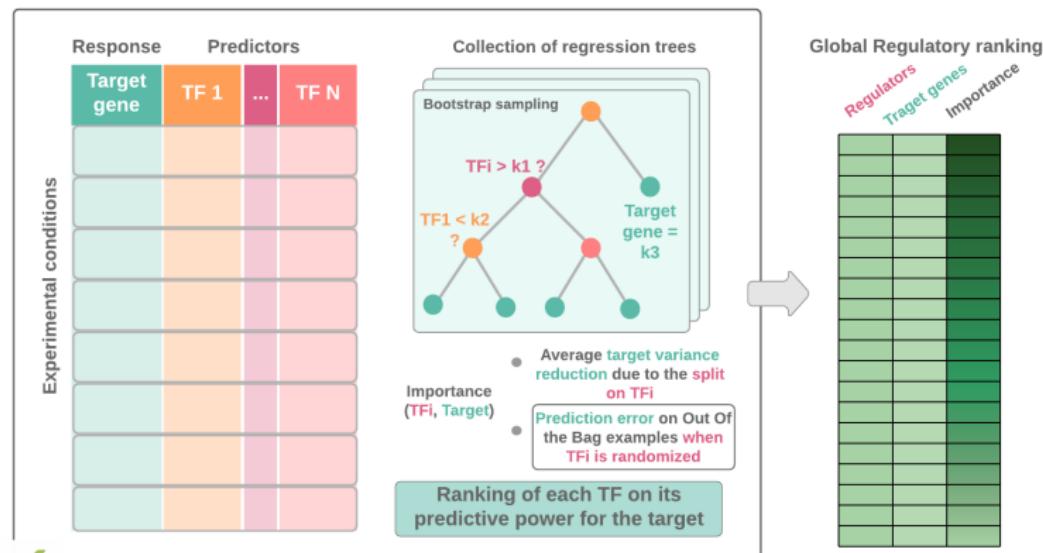




Exemple de méthodes : TIGRESS et GENIE3

GENIE3: étapes de la procédure

Ranking the regulators according to their relevance for predicting the other genes expression



Each gene in the input list becomes the target



Validation et perspectives

Principe de la régression pour les réseaux de régulation

regulators_i : niveaux d'expression des régulateurs transcriptionnels dans la condition i

target_i : niveaux d'expression d'un gène cible dans la condition i

$$\text{target}_i = \mathbf{f}(\text{regulators}_i) + \epsilon_i$$

La procédure de construction de réseau est la suivante :

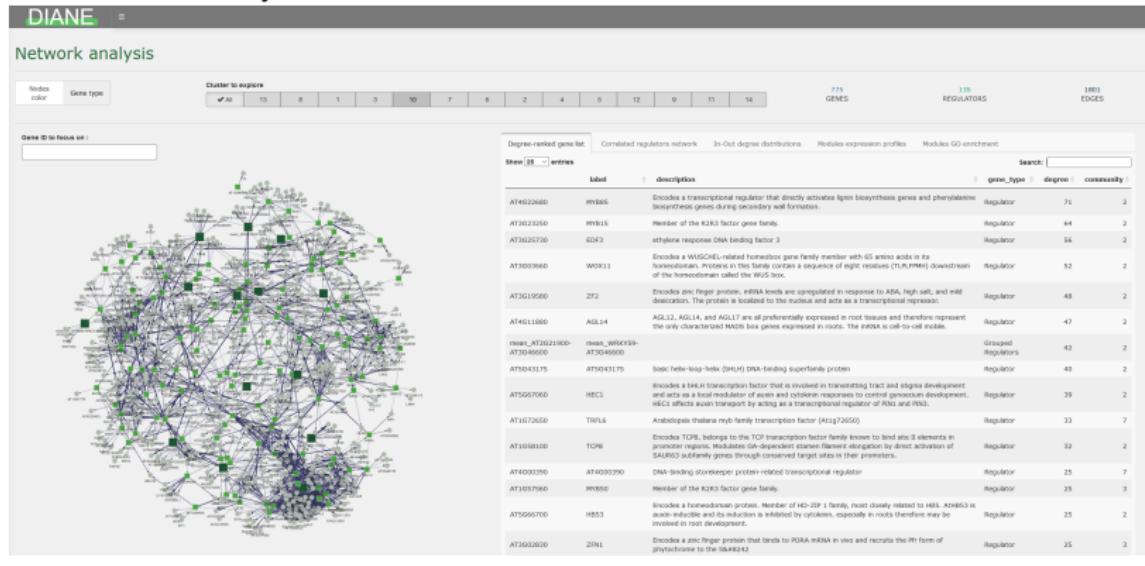
- ① Pour chaque gène du jeu de données, ajuster à partir des valeurs d'expression la fonction \mathbf{f}
- ② Extraire de \mathbf{f} les scores (ou valeur d'influence, importance, pouvoir prédictif) des régulateurs sur chaque gène du jeu de données
- ③ Sélectionner les scores régulateurs-gènes cibles les plus forts pour construire le réseau final

DIANE [Cassan et al., 2021]

Dashboard for the Inference and Analysis of Networks from Expression data



L'outil que vous utiliserez lors des TP-TD pour aller de données d'expression brutes jusqu'à l'inférence et l'analyses de réseau avec GENIE3.





Enrichir et valider un réseau inféré

Les arêtes inférées peuvent être comparées à des **liens de régulation déjà documentés**, comme :

- Les interactions présentes dans la **littérature**
- Des **données de fixation** des TFs *in vivo* ou *in vitro* CHIPSeq, DAPSeq
- L'**accessibilité de la chromatine** et footprinting : ATACSeq
- La **régulation in planta** (induction de TF dans des protoplastes [Bargmann et al., 2013], expression de gènes cibles dans des lignées de mutants, etc)



Enrichir et valider un réseau inféré

Les arêtes inférées peuvent être comparées à des **liens de régulation déjà documentés**, comme :

- Les interactions présentes dans la **littérature**
- Des **données de fixation** des TFs *in vivo* ou *in vitro* CHIPSeq, DAPSeq
- L'**accessibilité de la chromatine** et footprinting : ATACSeq
- La **régulation in planta** (induction de TF dans des protoplastes [Bargmann et al., 2013], expression de gènes cibles dans des lignées de mutants, etc)

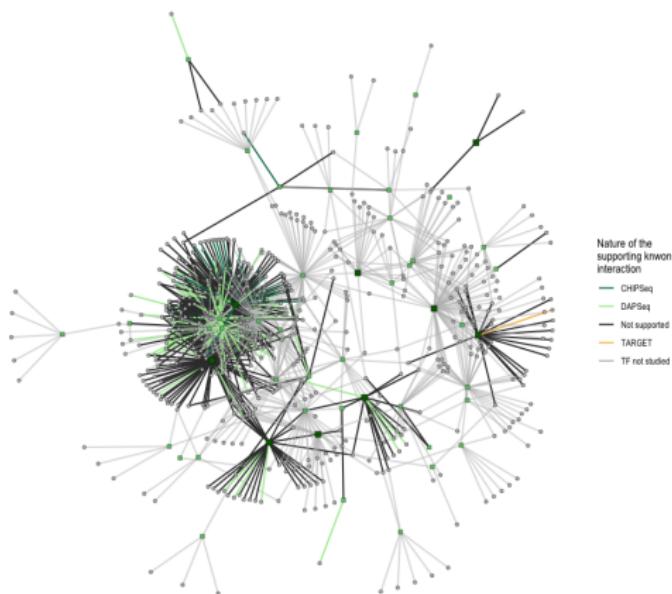
Quelques efforts de regroupement en bases de données:

- ConnecTF [Brooks et al., 2020] (Arabidopsis, maïs)
- AtRegNet [Palaniswamy et al., 2006] (Arabidopsis)

Enrichir et valider un réseau inféré

Network edges colored according to their experimental evidence

27.31 % of the edges (with validation information available) are supported



Ici, les arêtes d'un réseau prédit sont colorées suivant leur confirmation par une expérience présente dans connecTF (DAPSeq, CHIPSeq, TARGET)

Réseau inféré via GENIE3, validé via AraNetBench. Arabidopsis sous stress osmotique, salin, et en température

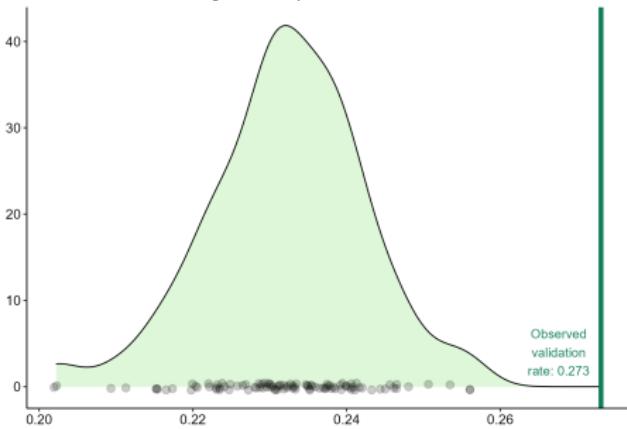
Calculer des métriques de validation sur un réseau inféré

- **Vrais positifs - précision :**
nombre d'arêtes prédites supportées par une information expérimentale (absolu, ou rapporté au nombre d'arêtes total qu'il est possible de valider)
- **Faux positifs, vrais négatifs, faux négatifs, rappel**

Calculer des métriques de validation sur un réseau inféré

- **Vrais positifs - précision :**
nombre d'arêtes prédites supportées par une information expérimentale (absolu, ou rapporté au nombre d'arêtes total qu'il est possible de valider)
- **Faux positifs, vrais négatifs, faux négatifs, rappel**

Testing the inferred network's validation rate : P value = 0 Z-score = 4.009
Null distribution of validated edges rates computed on 100 shuffled networks

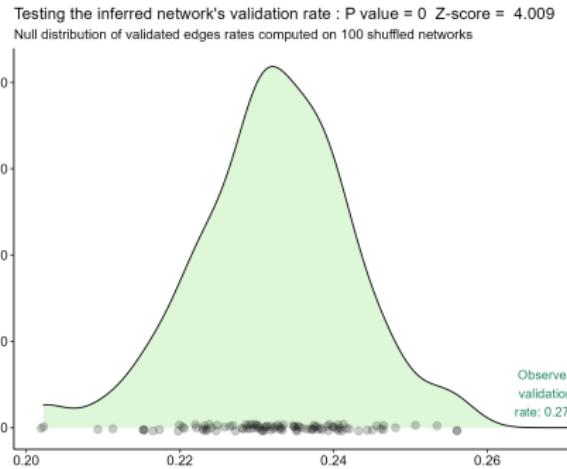


Calculer des métriques de validation sur un réseau inféré

- **Vrais positifs - précision :**
nombre d'arêtes prédites supportées par une information expérimentale (absolu, ou rapporté au nombre d'arêtes total qu'il est possible de valider)
- **Faux positifs, vrais négatifs, faux négatifs, rappel**

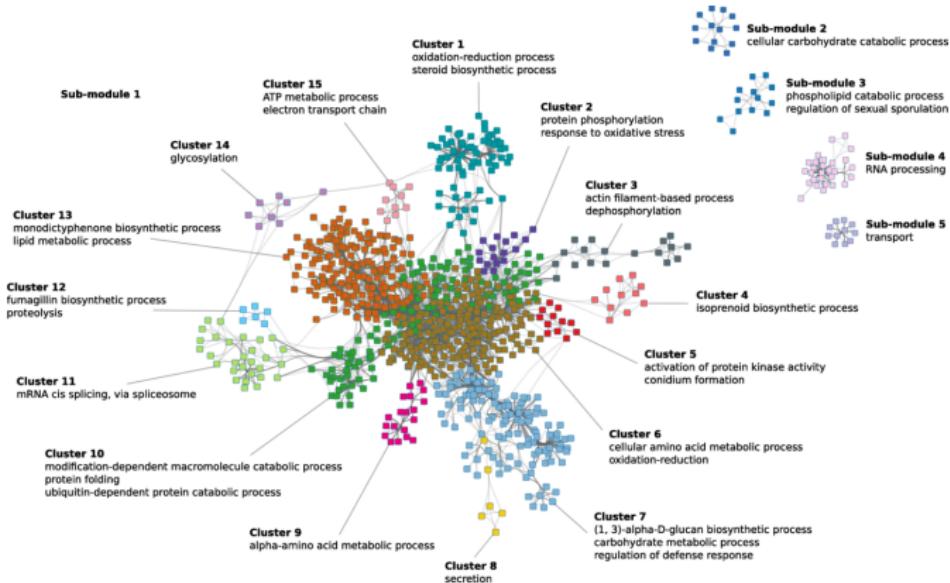
Interprétation de ces métriques

Ces données de validation sont **imparfaites**, elles contiennent des faux positifs, et faux négatifs : prudence



Analyser la topologie d'un réseau : détection de modules

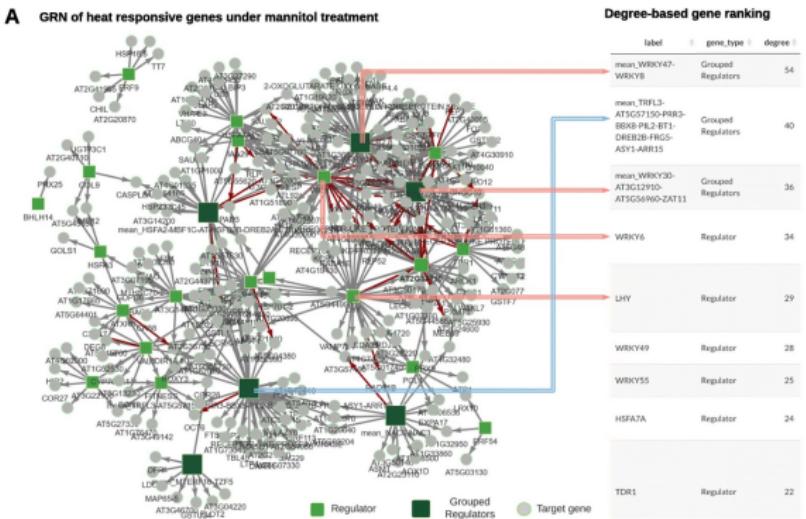
- **Communautés** de gènes densément connectés, contenant des éventuels enrichissements ontologiques. Conrad et al, BMC Syst. Biology 2018



Analyser la topologie d'un réseau inféré : connectivité

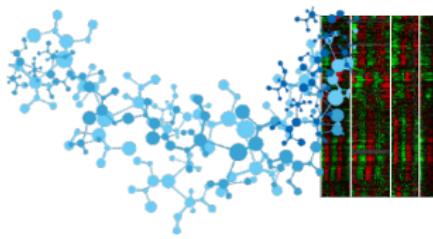
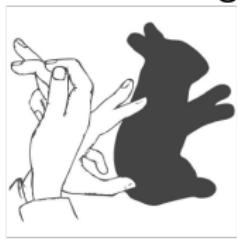
- **Degré - centralité** : Les gènes montrant une connectivité remarquable dans le réseau sont de potentiels régulateurs clés

From: Inferring and analyzing gene regulatory networks from multi-factorial expression data: a complete and interactive suite



Inférer des réseaux de régulation : un tâche encore complexe

① Problème en grande dimension

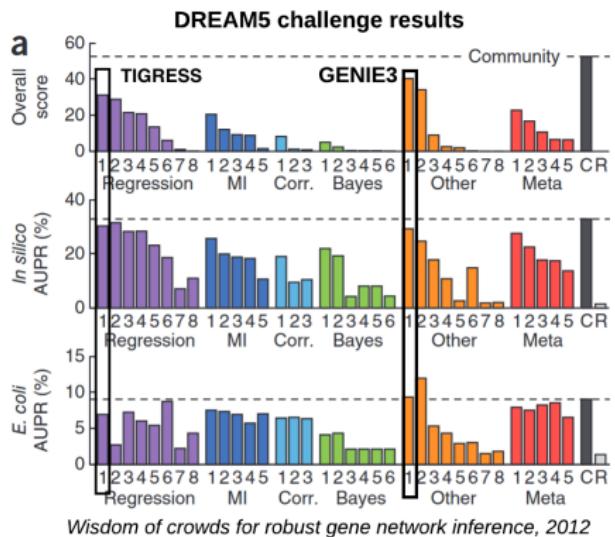


Q

② Manque de données de validation complètes et sûres pour étalonner les méthodes

Combiner plusieurs approches d'inférence

En 2012, les challenges **DREAM** ont évalué et combiné l'état de l'art des méthodes d'inférence et conclu à un apport significatif de la combinaison de plusieurs méthodes [Marbach et al., 2012]





Utilité de ces analyses pour la biologie des systèmes

L'analyse des réseaux inférés peut permettre de:

- Conforter et approfondir des connaissances existantes en biologie des systèmes, annoter de nouveaux gènes
- Découvrir de nouveaux gènes candidats contrôlant des réponses d'intérêt, après validation expérimentale et étude fonctionnelle
- Générer de nouvelles hypothèses et réduire l'espace de recherche pour les biologistes

Meilleure compréhension des systèmes vivants, solutions pour améliorer la résilience d'un organisme à une contrainte environnementale ou à des pathologies



References I

- ▶ Bargmann, B. O., Marshall-Colon, A., Efroni, I., Ruffel, S., Birnbaum, K. D., Coruzzi, G. M., and Krouk, G. (2013).
TARGET: A transient transformation system for genome-wide transcription factor target discovery.
Molecular Plant, 6(3):978–980.
- ▶ Brooks, M. D., Juang, C.-L., Katari, M. S., Alvarez, J. M., Pasquino, A., Shih, H.-J., Huang, J., Shanks, C., Cirrone, J., and Coruzzi, G. M. (2020).
ConnecTF: A platform to integrate transcription factor–gene interactions and validate regulatory networks.
Plant Physiology, 185(1):49–66.
- ▶ Cassan, O., Lèbre, S., and Martin, A. (2021).
Inferring and analyzing gene regulatory networks from multi-factorial expression data: a complete and interactive suite.
BMC Genomics, 22(1).
- ▶ 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(1):145.



References II

- ▶ 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.
- ▶ Palaniswamy, S. K., James, S., Sun, H., Lamb, R. S., Davuluri, R. V., and Grotewold, E. (2006).
AGRIS and AtRegNet. a platform to link cis-regulatory elements and transcription factors into regulatory networks.
Plant Physiology, 140(3):818–829.
- ▶ Sanguinetti, G. and Huynh-Thu, V. A. (2019).
Gene regulatory networks.
Book edited by Springer.