Méthodes statistiques pour l'inférence de réseaux de régulation

Océane Cassan, Sophie Lèbre
oceane.cassan@lirmm.fr
sophie.lebre@umontpellier.fr

BFP M1, Parcours Bipa

2 Février 2024

Introduction

Inférence de GRN par régression

2 exemples de méthodes

Régression linéaire ou non

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

$$target_i = \mathbf{f}(regulators_i) + \epsilon_i$$

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]

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 c**ombinaison linéaire** de l'expression des facteurs de transcription :

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 c**ombinaison linéaire** de l'expression des facteurs de transcription :

$$target_i = \beta_{target,1}.TF1_i + \beta_{target,2}.TF2_i + ... + \beta_{target,M}.TFM_i + \epsilon_i$$

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

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 :

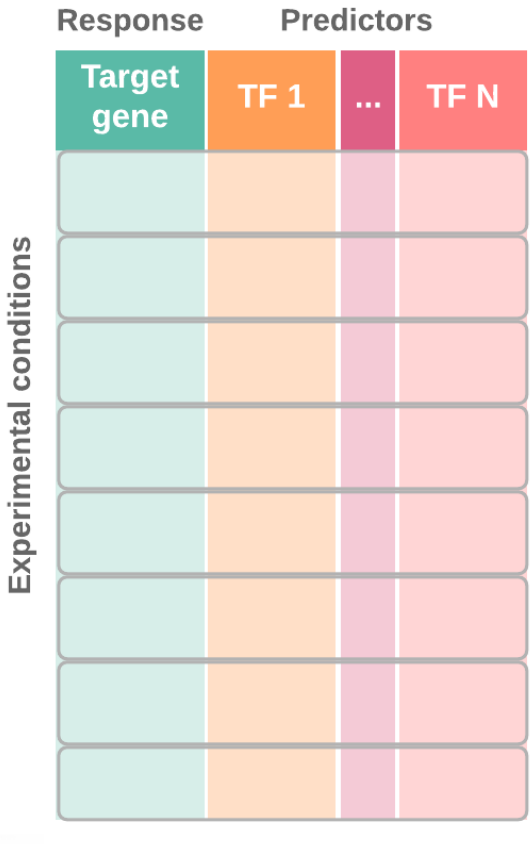
$$target_i = \beta_{target,1}.TF1_i + \beta_{target,2}.TF2_i + ... + \beta_{target,M}.TFM_i + \epsilon_i$$

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

Problème : il faut refléter la sparsité du problème biologique

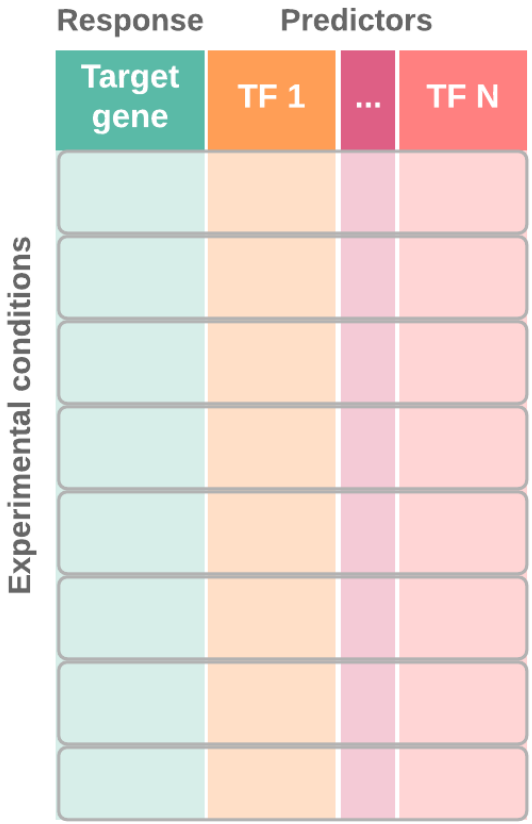
L'expression d'un gène est sensée être expliquée par un nombre limité de TFs, et non tous les TFs du jeu de données → **LARS** (Least-angle regression)

TIGRESS : étapes de la procédure



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

TIGRESS : étapes de la procédure

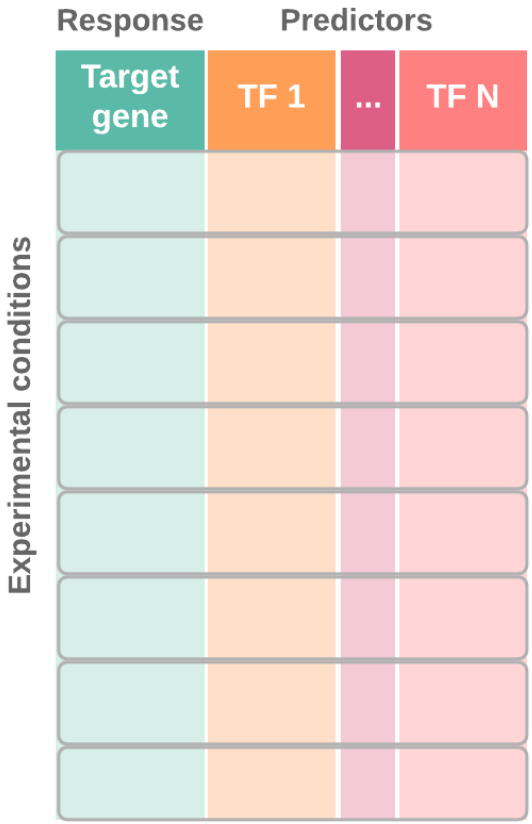


LARS : add L predictors iteratively, each time choosing which one helps to explains the most X_{target}

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

- Top L predictors: TF_2 ... TF_8 ... TF_5
- N-L TFs not selected

TIGRESS : étapes de la procédure



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

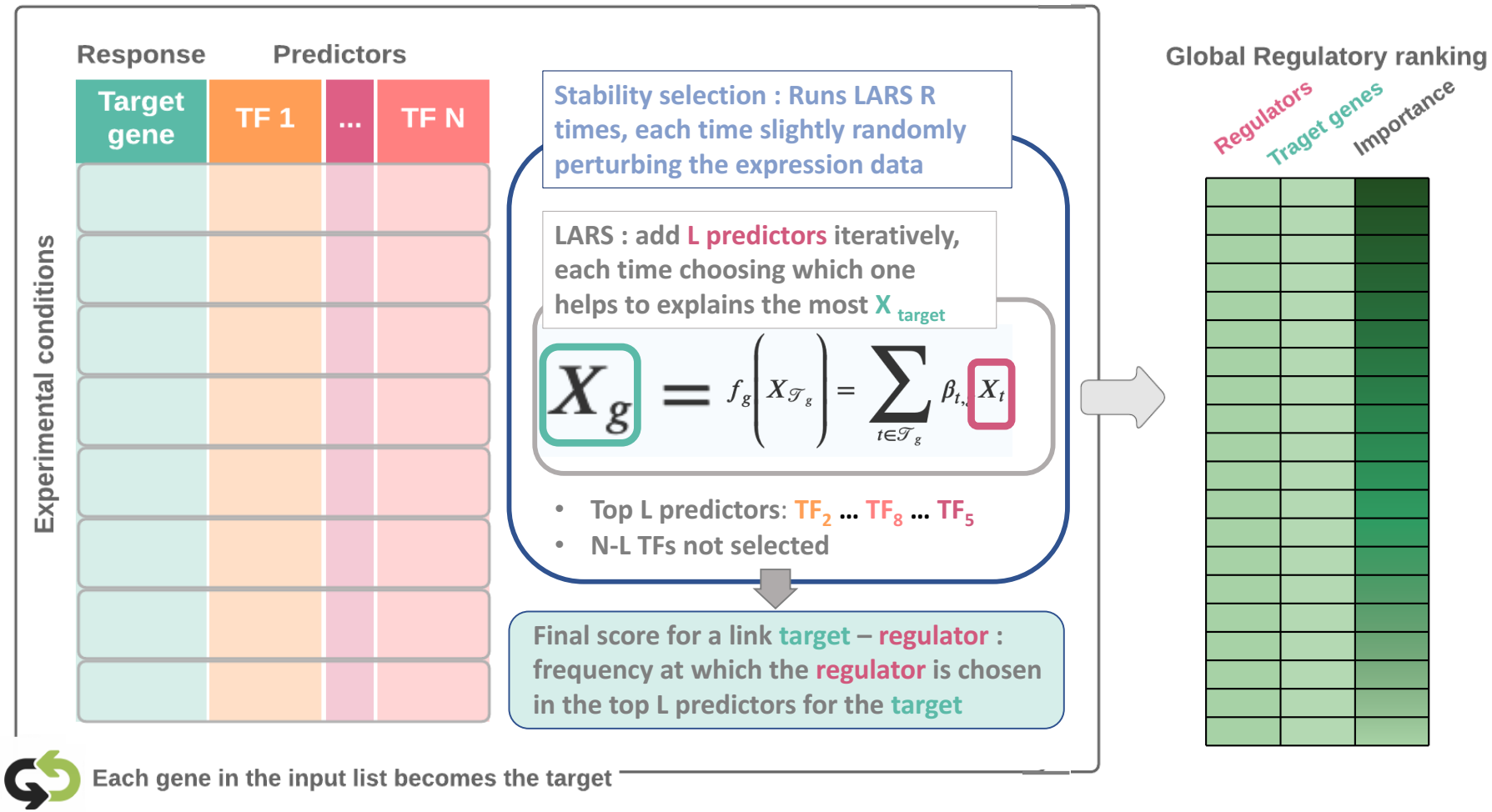
LARS : add L predictors iteratively, each time choosing which one helps to explains the most X_{target}

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

- Top L predictors: TF_2 ... TF_8 ... TF_5
- 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

TIGRESS : étapes de la procédure



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 :

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 :

$$target_i = RandomForest(TF_i) + \epsilon_i$$

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

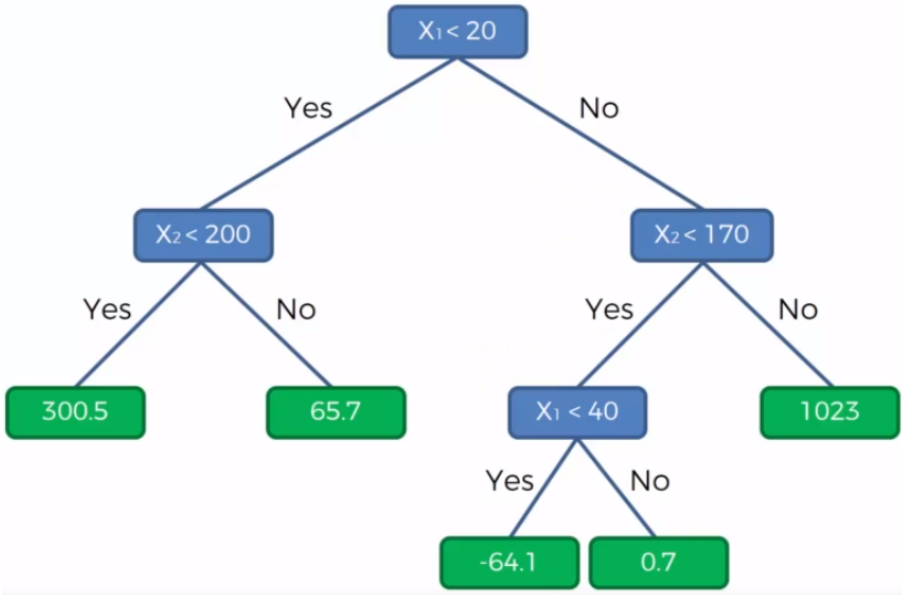
Avec TF_i le niveau d'expression de tous les TFs du jeu de données dans la condition i.

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

GENIE3 : approche basée sur les arbres de régression

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

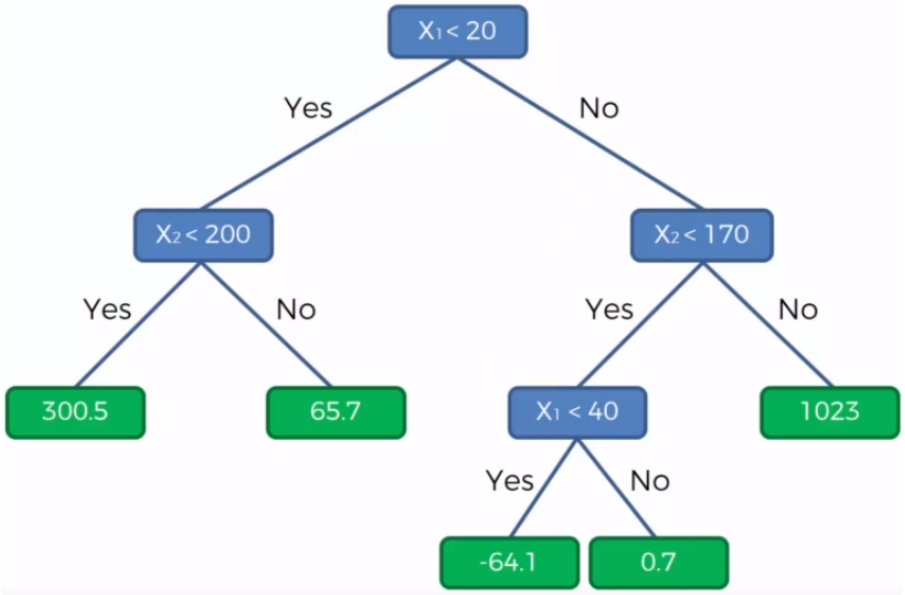


Ajustement d'un arbre de régression

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

GENIE3 : approche basée sur les arbres de régression

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



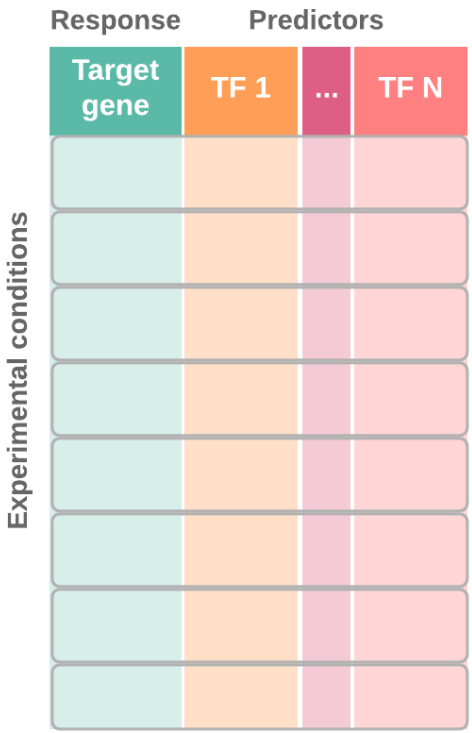
Ajustement d'un arbre de régression

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 échantillonnées légèrement différemment les unes des autres → leur consensus permet plus de robustesse dans les prédictions (apprentissage ensembliste)

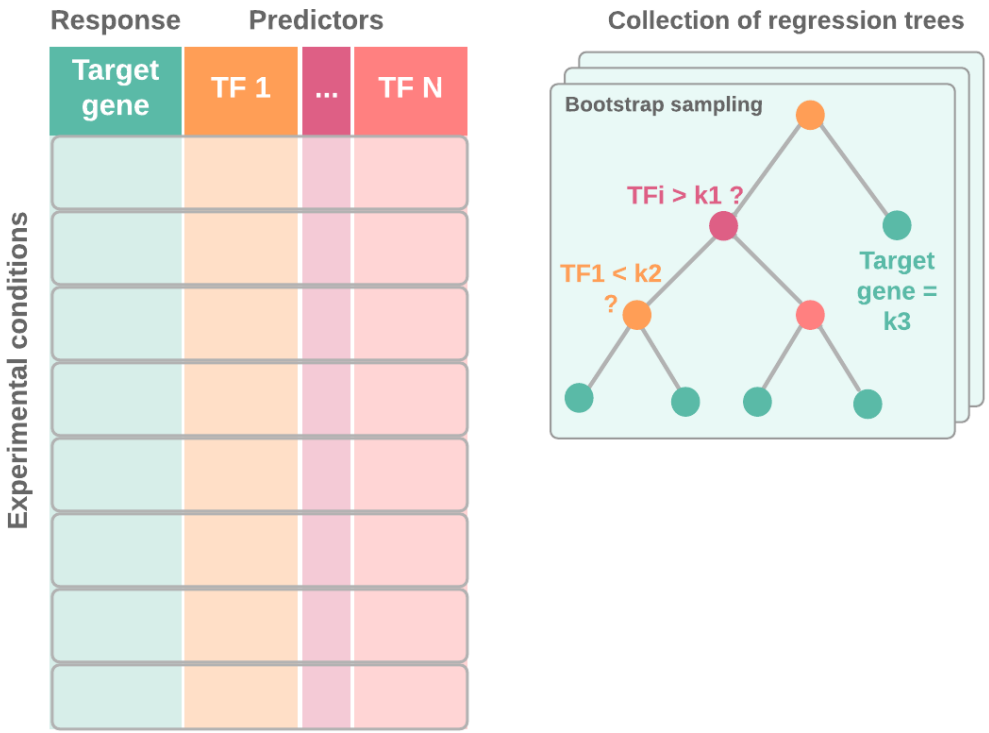
GENIE3: étapes de la procédure

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



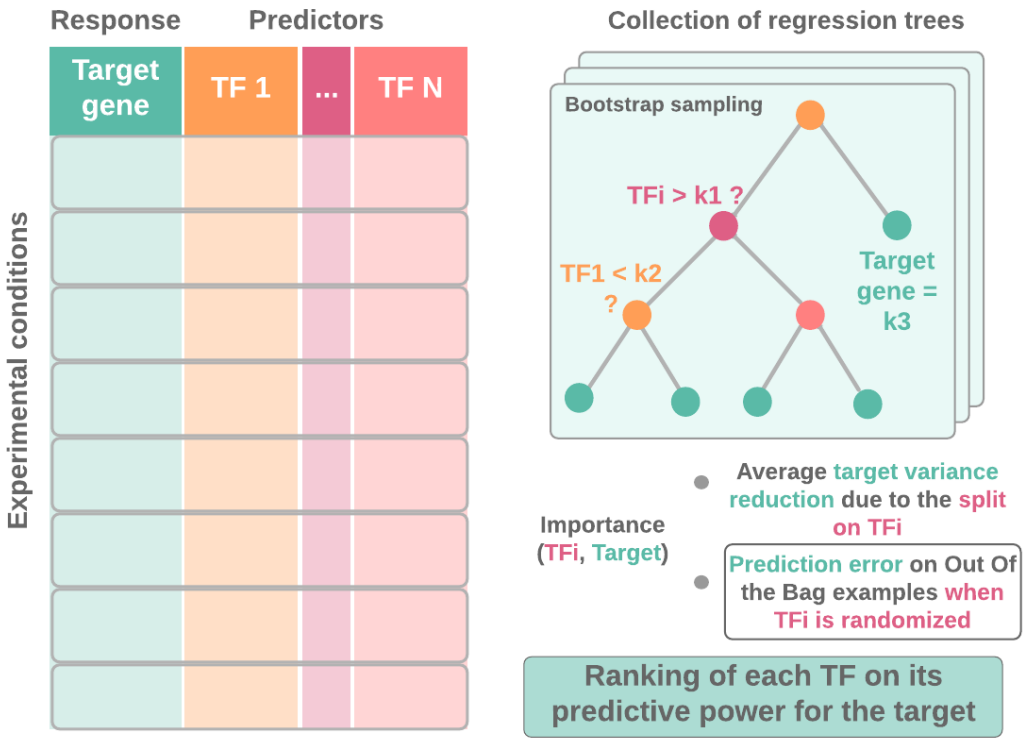
GENIE3: étapes de la procédure

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



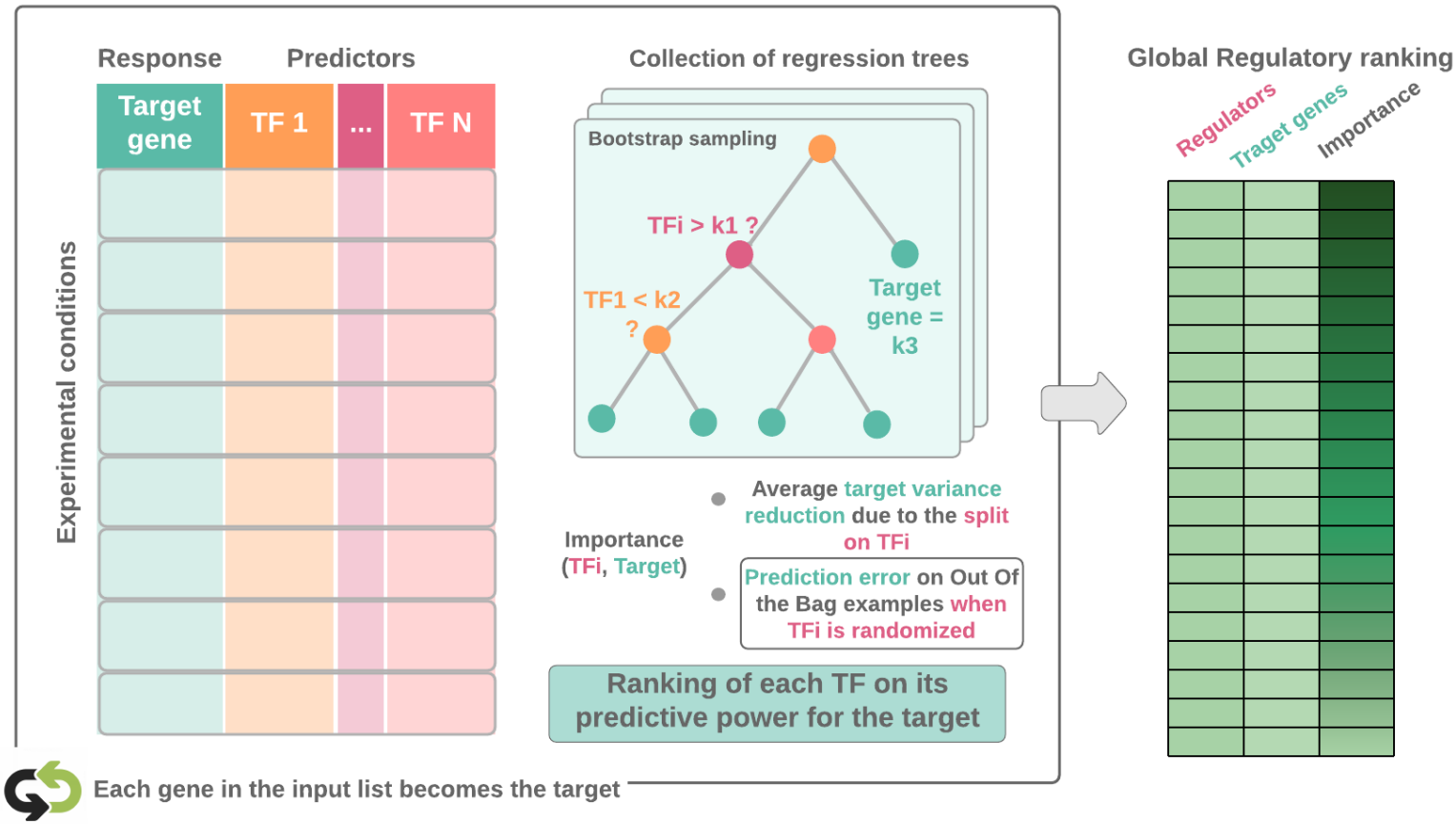
GENIE3: étapes de la procédure

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



GENIE3: étapes de la procédure

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



Validation et perspectives

Principe de la régression pour les réseaux de régulation

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

$$target_i = \mathbf{f}(regulators_i) + \epsilon_i$$

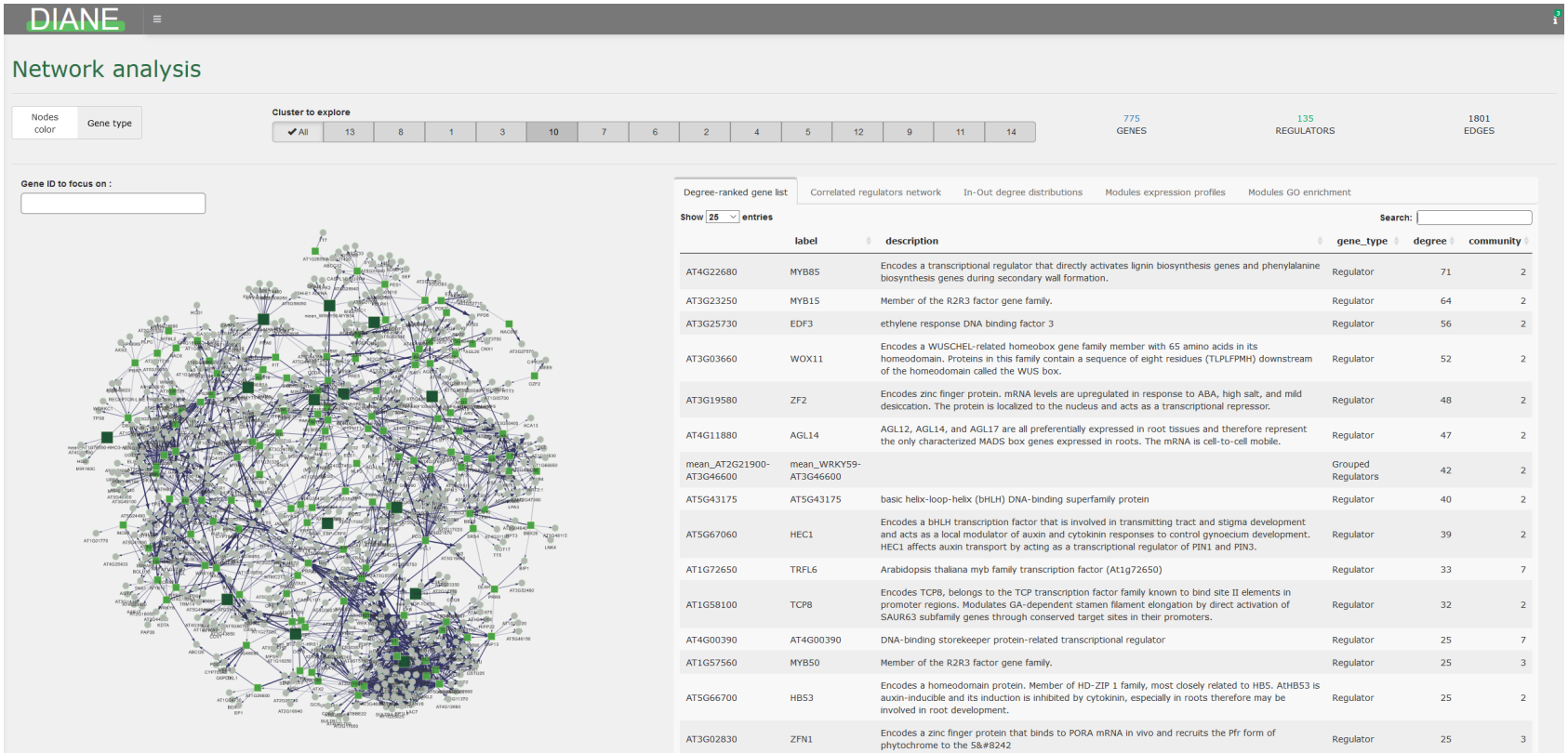
La procédure de construction de réseau est la suivante :

1. Pour chaque gène du jeu de données, ajuster à partir des valeurs d'expression la fonction **f**
2. Extraire de **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.



	label	description	gene_type	degree	community
AT4G22680	MYB85	Encodes a transcriptional regulator that directly activates lignin biosynthesis genes and phenylalanine biosynthesis genes during secondary wall formation.	Regulator	71	2
AT3G23250	MYB15	Member of the R2R3 factor gene family.	Regulator	64	2
AT3G25730	EDF3	ethylene response DNA binding factor 3	Regulator	56	2
AT3G03660	WOX11	Encodes a WUSCHEL-related homeobox gene family member with 65 amino acids in its homeodomain. Proteins in this family contain a sequence of eight residues (TLPLFPMH) downstream of the homeodomain called the WUS box.	Regulator	52	2
AT3G19580	ZF2	Encodes zinc finger protein. mRNA levels are upregulated in response to ABA, high salt, and mild desiccation. The protein is localized to the nucleus and acts as a transcriptional repressor.	Regulator	48	2
AT4G11880	AGL14	AGL12, AGL14, and AGL17 are all preferentially expressed in root tissues and therefore represent the only characterized MADS box genes expressed in roots. The mRNA is cell-to-cell mobile.	Regulator	47	2
mean_AT2G21900-AT3G46600	mean_WRKY59-AT3G46600		Grouped Regulators	42	2
AT5G43175	AT5G43175	basic helix-loop-helix (bHLH) DNA-binding superfamily protein	Regulator	40	2
AT5G67060	HEC1	Encodes a bHLH transcription factor that is involved in transmitting tract and stigma development and acts as a local modulator of auxin and cytokinin responses to control gynoecium development. HEC1 affects auxin transport by acting as a transcriptional regulator of PIN1 and PIN3.	Regulator	39	2
AT1G72650	TRFL6	Arabidopsis thaliana myb family transcription factor (At1g72650)	Regulator	33	7
AT1G58100	TCP8	Encodes TCP8, belongs to the TCP transcription factor family known to bind site II elements in promoter regions. Modulates GA-dependent stamen filament elongation by direct activation of SAUR63 subfamily genes through conserved target sites in their promoters.	Regulator	32	2
AT4G00390	AT4G00390	DNA-binding storekeeper protein-related transcriptional regulator	Regulator	25	7
AT1G57560	MYB50	Member of the R2R3 factor gene family.	Regulator	25	3
AT5G66700	HB53	Encodes a homeodomain protein. Member of HD-ZIP 1 family, most closely related to HB5. AtHB53 is auxin-inducible and its induction is inhibited by cytokinin, especially in roots therefore may be involved in root development.	Regulator	25	2
AT3G02830	ZFN1	Encodes a zinc finger protein that binds to PORA mRNA in vivo and recruits the Pfr form of phytochrome to the 5&#8242	Regulator	25	3

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 **litérature**
- 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

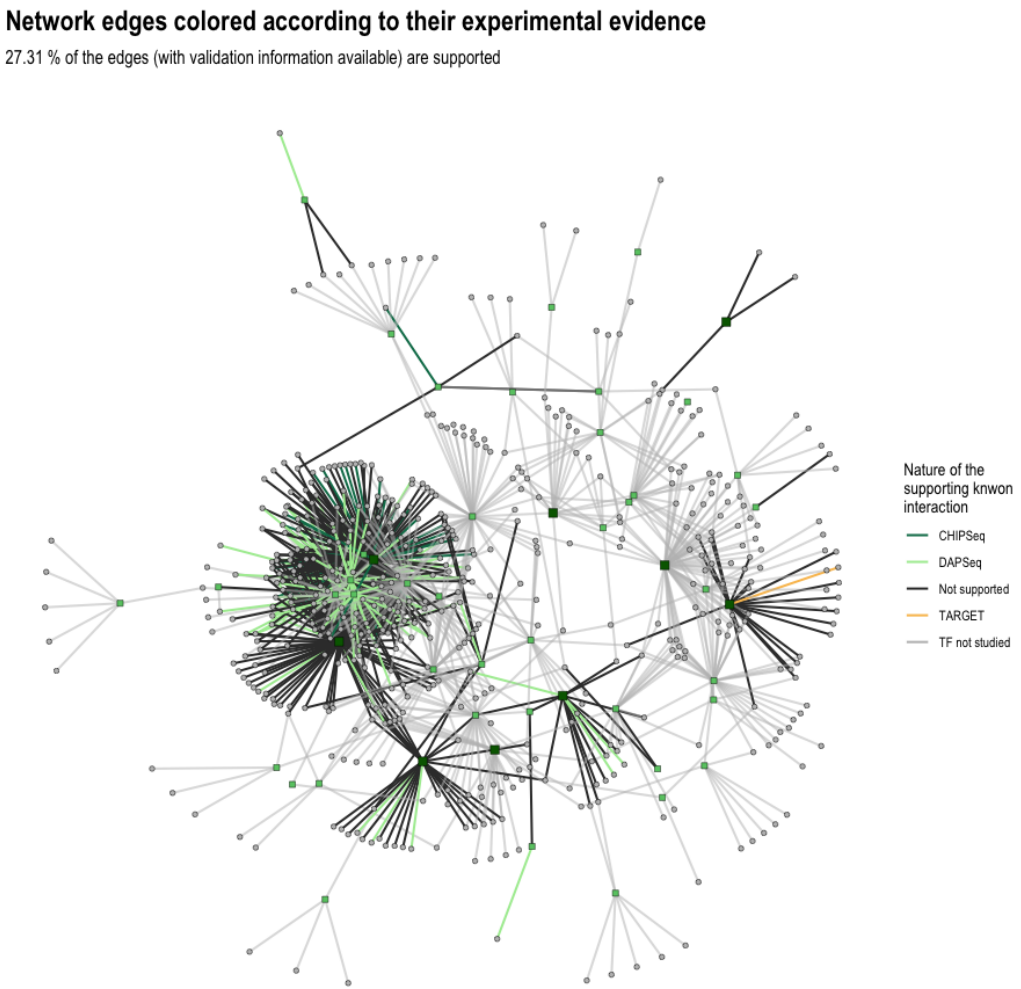
Les arêtes inférées peuvent être comparées à des **liens de régulation déjà documentés**, comme :

- Les interactions présentes dans la **litérature**
- 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



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

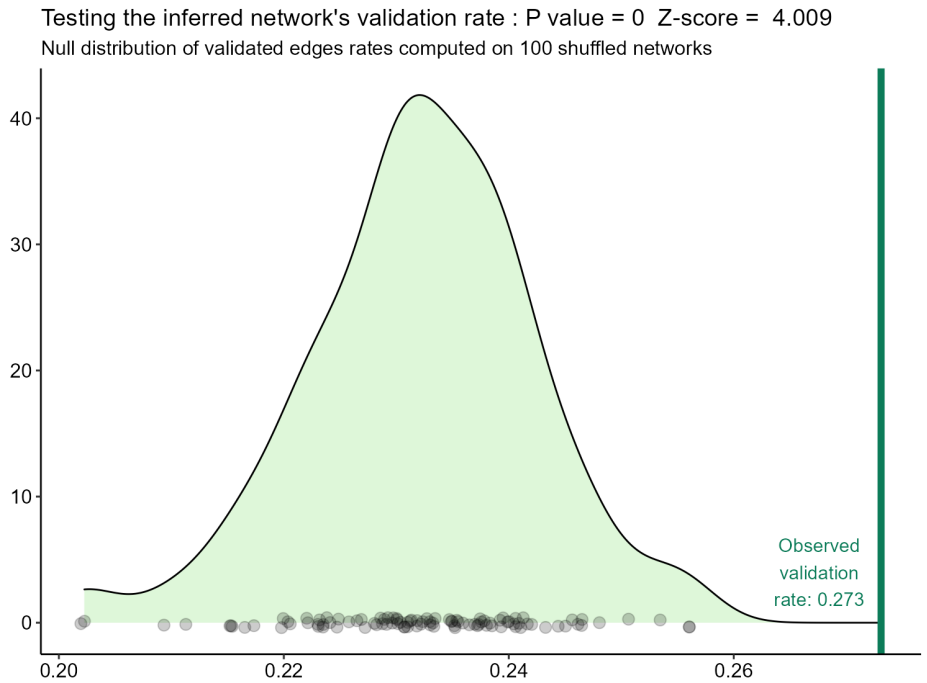
Réseau inféré via GENIE3, validé via AraNetBench. Arabidopsis sous stress osmotique, salin, et en température

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**

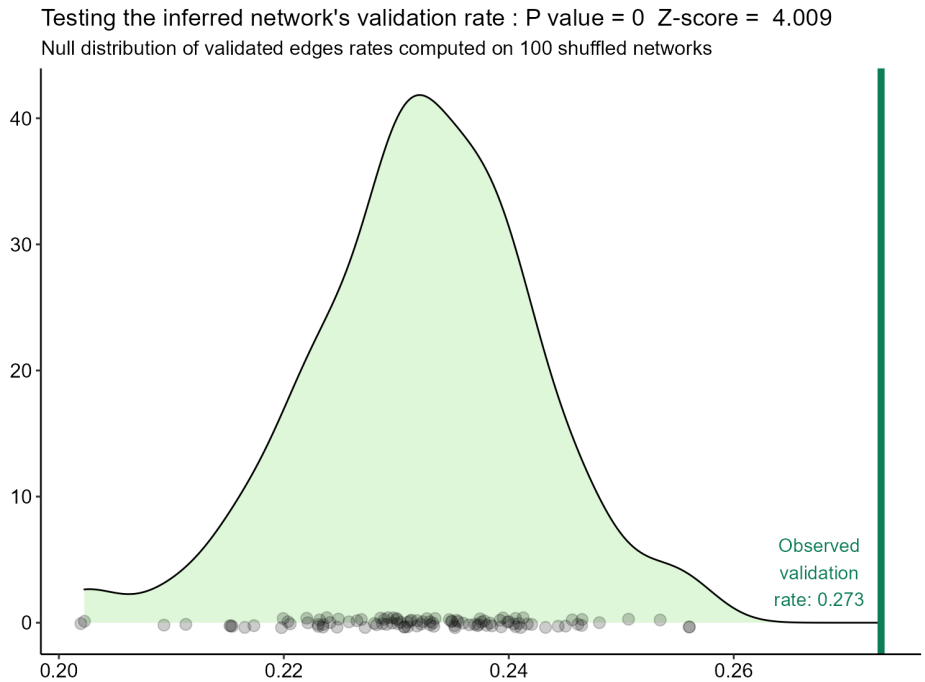
Calculer des métriques de validation sur un réseau inféré

- **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**



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**

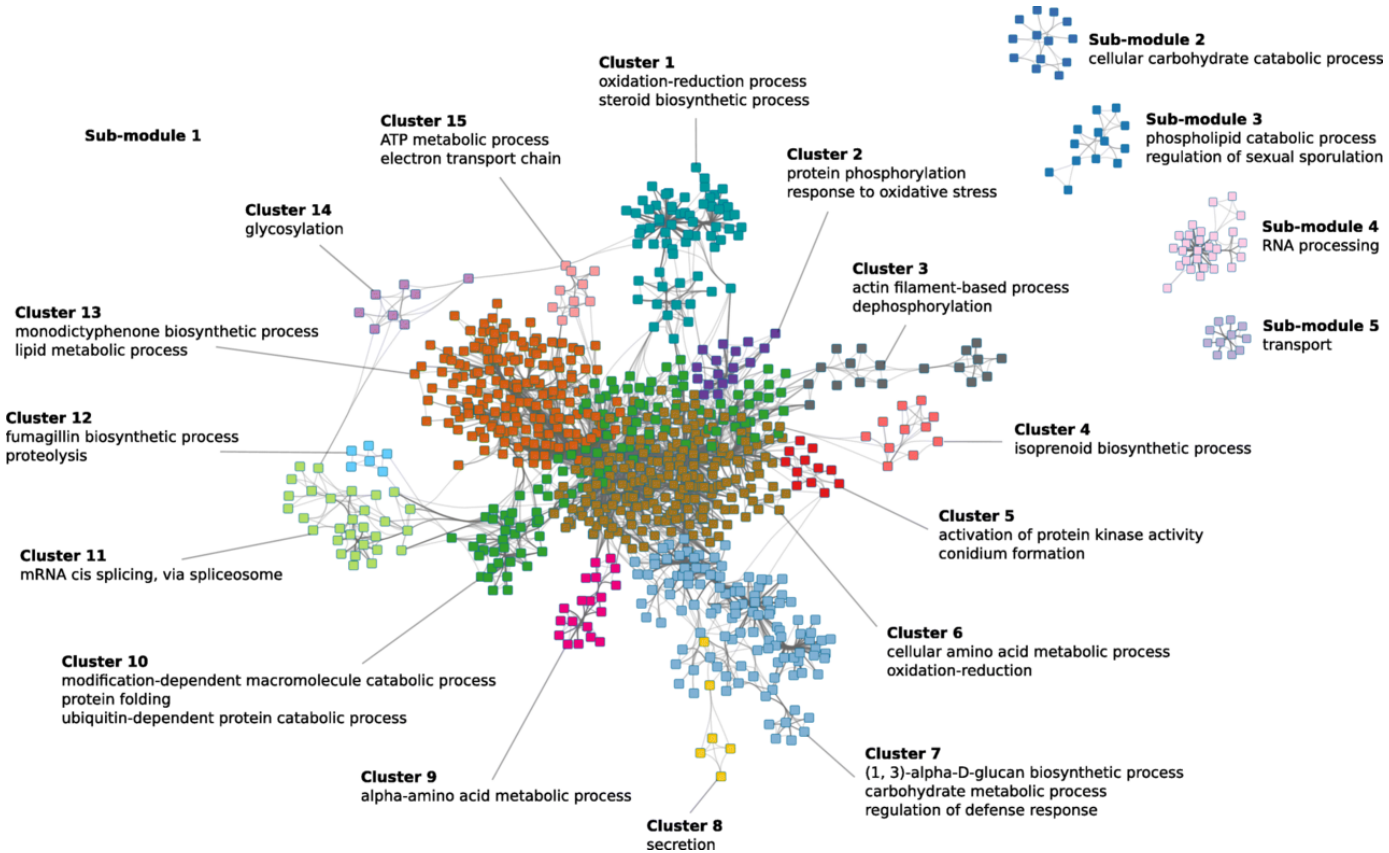


Interprétation de ces métriques

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

Analyser la topologie d'un réseau : détection de modules

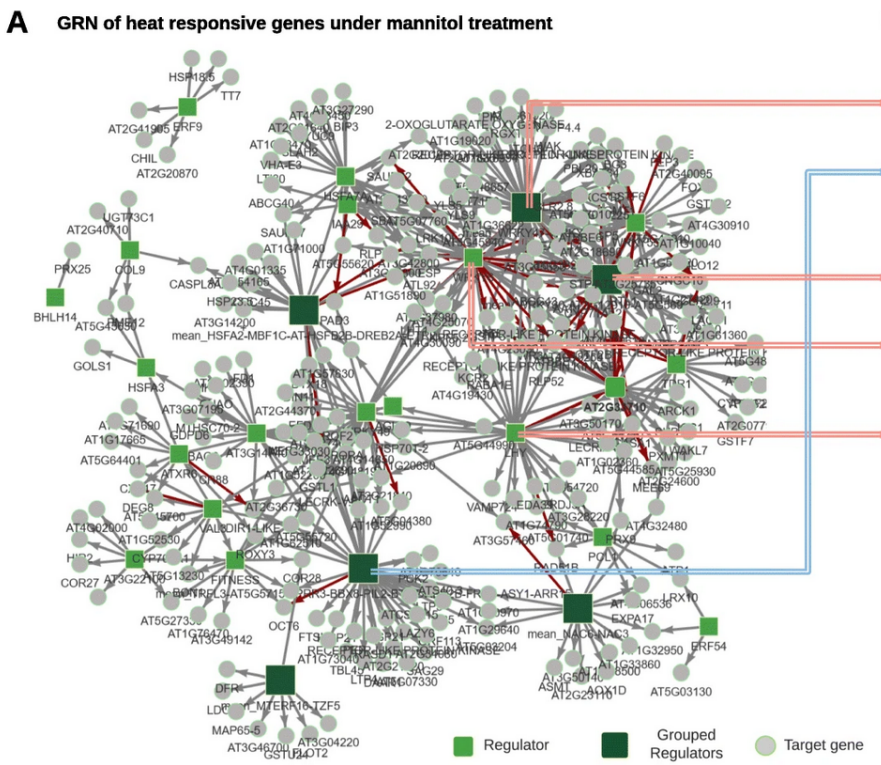
- **Communautés** de gènes densément connectés, contenant des éventuels enrichissements ontologiques. Conrad et al, BMC Syst. Biology 2018



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

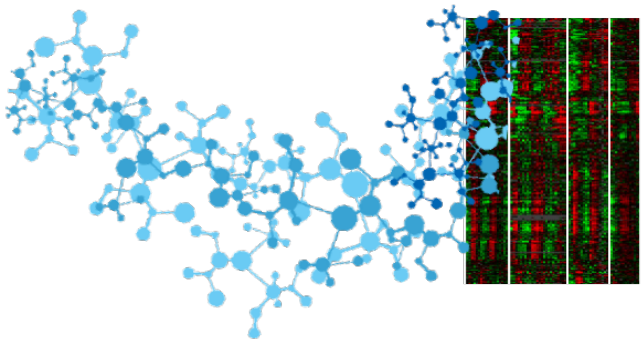


Degree-based gene ranking

label	gene_type	degree
mean_WRKY47-WRKY8	Grouped Regulators	54
mean_TRFL3-AT5G57150-PRR3-BBX8-PIL2-BT1-DREB2B-FRG5-ASY1-ARR15	Grouped Regulators	40
mean_WRKY30-AT3G12910-AT5G56960-ZAT11	Grouped Regulators	36
WRKY6	Regulator	34
LHY	Regulator	29
WRKY49	Regulator	28
WRKY55	Regulator	25
HSFA7A	Regulator	24
TDR1	Regulator	22

Inférer des réseaux de régulation : un tâche encore complexe

1. Problème en grande dimension

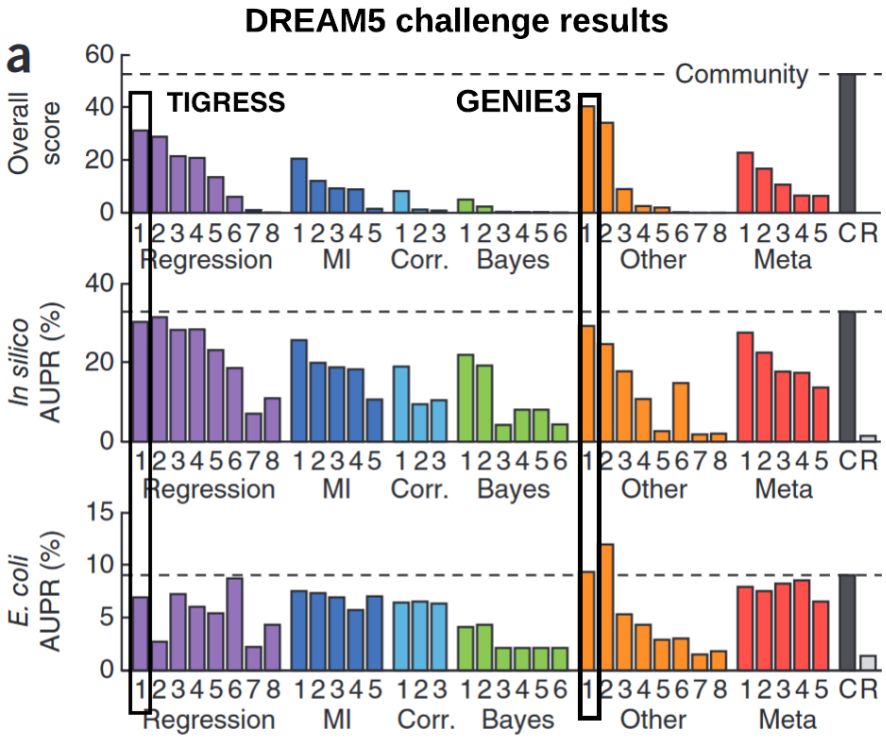


Q

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]



Wisdom of crowds for robust gene network inference, 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.