Classification et inférence de réseaux de gènes à partir de séries temporelles très courtes : application à la modélisation de la mémoire transcriptionnelle végétale associée à des stimulations sonores répétées

Khaoula Hadj Amor

RÉPUBLIQUE FRANÇAISE
Liberté Égalité Fraternité

Université de Toulouse

Doctorat de l'Université de Toulouse

préparé à l'Université Toulouse III - Paul Sabatier

Classification et inférence de réseaux de gènes à partir de séries temporelles très courtes. Application à la modélisation de la mémoire transcriptionnelle végétale associée à des stimulations sonores répétées.

Thèse présentée et soutenue, le 21 mars 2024 par

Khaoula HADJ AMOR

École doctorale
EDMITT - Ecole Doctorale Mathématiques, Informatique et Télécommunications de Toulouse

Spécialité
Mathématiques et Applications

Unité de recherche
MIAT - Mathématiques et Informatique Appliquées Toulouse

Thèse dirigée par
Frédérick GARCIA et Adelin BARBACCI

Composition du jury
M. Jean-Philippe GALAUD, Président, Université Toulouse III - Paul Sabatier
M. Jean-Louis JULIEN, Rapporteur, Université Clermont Auvergne
Mme Marie-Laure MARTIN-MAGNIETTE, Rapporteure, INRAE Ile-de-France - Versailles-Saclay
M. Harold DURUFLÉ, Examinateur, INRAE Val de Loire
Mme Sophie LÈBRE, Examinatrice, Université Paul Valéry Montpellier 3
M. Frédérick GARCIA, Directeur de thèse, INRAE Occitanie-Toulouse
M. Adelin BARBACCI, Co-directeur de thèse, INRAE Occitanie-Toulouse

Remerciements

Je tiens tout d'abord à remercier mes deux encadrants, mon directeur de thèse, Frédérick Garcia, et mon co-directeur de thèse, Adelin Barbacci. Je vous suis reconnaissante pour votre encadrement de qualité et vos précieux conseils. Merci, Frédérick, d'avoir toujours trouvé du temps pour moi, ta rigueur et ton dévouement envers la recherche ont été une source constante d'encouragement, m'incitant à m'investir davantage dans mon travail. Merci, Adelin, de m'avoir initiée à la biologie et à la biomécanique, et de m'avoir aidée à simplifier mes concepts. Ta disponibilité permanente ainsi que tes mots d'encouragement et tes conseils pendant les moments difficiles ont été d'une grande valeur pour moi. Cela a été un réel plaisir d'être encadrée par vous.

Je remercie également la région Occitanie et l'INRAE pour le financement de mes deux premières années de thèse, ainsi que l'Université de Toulouse 3 pour le financement de mes deux dernières années de thèse en tant qu'ATER.

Je remercie Marie-Laure Martin et Jean-Louis Julien d'avoir accepté de consacrer une partie de leur temps à la lecture de ce manuscrit. Je remercie également Jean-Philippe Galaud, Sophie Lèbre et Harold Duruflé d'avoir accepté de faire partie de mon jury de thèse.

Je souhaite exprimer ma gratitude à l'unité MIAT, avec un remerciement spécial au directeur d'unité, Sylvain Jasson. Cette thèse m'a permis de rencontrer des personnes extraordinaires. Merci pour nos discussions enrichissantes et pour les moments agréables passés ensemble à la salle café.

Je tiens à exprimer ma gratitude également envers l'équipe QIP de LIPMe, en

particulier envers le chef d'équipe, Sylvain Raffaele. Sa bienveillance a été toujours présente lors de nos discussions. Même si je ne comprenais pas tout lors des réunions d'équipe, y participer a été un réel plaisir pour moi.

Merci à tous les doctorants du labo (MIAT et QIP) pour les moments et discussions non-scientifiques partagés. Merci également pour les soirées jeux de société et les parties de tarot au labo. Je tiens à remercier en particulier mes co-bureaux, Pierre (Dr Pierre maintenant) et Aurélie, pour tous les moments que nous avons partagés ensemble. Merci de m'avoir écoutée pendant les périodes difficiles.

Mes remerciements vont également à toutes les personnes autour de moi, avec leurs questions récurrentes : "Quand prévois-tu de soutenir ta thèse ?", "Pourquoi n'as-tu pas encore publié ?", ou encore "Où en es-tu dans la rédaction ?". Bien qu'angoissantes et stressantes en périodes pleines de doutes, ces questions m'ont permis de maintenir le cap et de garder le focus sur mon objectif final.

Je souhaite également exprimer ma gratitude envers mes amis en dehors du laboratoire, ceux qui ne sont pas dans le domaine de la science et de la recherche. Merci d'avoir été toujours présents pour moi, de m'avoir écoutée même lorsque je n'avais pas beaucoup à dire en dehors de ma thèse. Nos sorties ensemble et les activités sportives que nous avons partagées ont été d'une grande aide pour moi pendant les périodes de stress, d'angoisse et de manque de sommeil.

Enfin, je tiens à exprimer toute ma gratitude envers ma famille. À mes parents, Mohamed et Najia, pour leur soutien inébranlable tout au long de cette aventure académique. Leur amour, leur encouragement constant et leur compréhension pendant les moments de stress ont été des piliers essentiels dans la réalisation de ce travail. Je leur suis infiniment reconnaissante pour leur présence, leur soutien moral et leur patience tout au long de cette thèse. À mes deux sœurs, Nawres et Rania, pour leur soutien constant, leurs paroles encourageantes et leur présence réconfortante qui ont été une source de motivation pendant cette période exigeante. Cette thèse est dédiée à 100% pour vous. Je vous aime !

Résumé

Les avancées des nouvelles technologies de séquençage ont ouvert l'accès aux données d'expression dynamique des gènes à l'échelle du génome. Les approches ensemblistes classiques, habituellement utilisées en biologie, ne permettent pas la compréhension des mécanismes moléculaires complexes sous-jacents. Par conséquent, le développement de méthodes analytiques permettant d'appréhender de manière plus satisfaisante ce type de données représente un défi majeur pour la biologie contemporaine. Cependant, les coûts techniques et expérimentaux associés aux données de transcriptomiques limitent la dimension des jeux de données réels et, par conséquent, leurs méthodes d'analyse.

Au cours de ma thèse, à l'interface entre les mathématiques appliquées et la biologie végétale, j'ai travaillé sur la mise en place d'une méthode d'inférence de réseaux de régulations dynamiques adaptée à un jeu de données réelles et originales décrivant l'effet de stimulations sonores répétées sur l'expression des gènes d'*Arabidopsis thaliana*. J'ai ainsi proposé une méthode de classification adaptée aux séries temporelles très courtes qui regroupe les gènes par variations temporelles, permettant d'ajuster la dimension des données à l'inférence de réseau. La comparaison de cette méthode aux méthodes classiques a permis de montrer qu'elle était la plus adaptée aux séries temporelles très courtes avec un pas de temps irrégulier. Pour l'inférence de réseau dynamique, j'ai proposé un modèle qui prend en compte la variabilité intra-classe et qui intègre un terme constant décrivant explicitement la stimulation externe du système. L'évaluation de ces méthodes de classification et d'inférence a été effectuée sur des données de séries temporelles simulées et réelles, ce qui a permis d'établir la bonne qualité des performance en terme de précision, de rappel et d'erreur de prédiction.

L'implémentation de ces méthodes a permis d'étudier le priming de la réponse immunitaire d'*Arabidopsis thaliana* par des ondes sonores répétées. Nous avons

montré l'existence de la formation d'une mémoire transcriptionnelle associée aux stimulations qui fait passer la plante d'un état naïf à un état primé et plus résistant en 3 jours. Cet état résistant, entretenu d'une part par les stimulations et d'autre part par des cascades de facteurs de transcription, augmente la résistance immunitaire de la plante en déclenchant l'expression de gènes de résistance chez la plante saine, en diversifiant le nombre de gènes participant à la réponse immunitaire et en intensifiant l'expression de nombreux gènes de résistance. L'inférence du réseau décrivant la mémoire transcriptionnelle associée aux stimulations sonores répétées nous a permis d'identifier les propriétés qu'elle confère à la plante. Ces prédictions, validées expérimentalement, ont montré par exemple que l'augmentation de la cadence entre stimulations ne permettait pas d'obtenir un gain de résistance plus conséquent et que la mémoire transcriptionnelle ne dure que 1.5 jours après la dernière stimulation.

Abstract

Advancements in new sequencing technologies have paved the way for accessing dynamic gene expression data on a genome-wide scale. Classical ensemble approaches traditionally used in biology fall short of comprehending the underlying the complex molecular mechanisms. Consequently, developing analytical methods to understand further such data poses a significant challenge for current biology. However, the technical and experimental costs associated with transcriptomic data severely limit the dimension of real datasets and their analytical methods.
Throughout my thesis, at the intersection of applied mathematics and plant biology, I focused on implementing an inference method for dynamic regulatory networks tailored to a real and original dataset describing the effect of repeated acoustic stimulations on genes expressions of *Arabidopsis thaliana*. I proposed a clustering method adapted to very-short time series that groups genes based on temporal variations, adjusting the data dimension for network inference. The comparison of this method with classical methods showed that it was the most suitable for very-short time series with irregular time points. For the network inference, I proposed a model that takes into account intra-class variability and integrates a constant term explicitly describing the external stimulation of the system. The evaluation of these classification and inference methods was conducted on simulated and real-time series data, which established their high performance in terms of accuracy, recall, and prediction error.
The implementation of these methods to study the priming of the immune response of *Arabidopsis thaliana* through repeated sound waves. We demonstrated the formation of a transcriptional memory associated with stimulations, transitioning the plant from a naïve state to a primed and more resistant state within 3 days. This resistant state, maintained by stimulations and transcription factor cascades, enhances the plant's immune resistance by triggering the expression of resistance

genes in healthy plants, diversifying the number of genes involved in the immune response, and intensifying the expression of numerous resistance genes. The inference of the network describing the transcriptional memory associated with repeated sound stimulations allowed us to identify the properties conferred to plants. Experimentally validated predictions showed that increasing the frequency between stimulations does not result in a more significant resistance gain, and the transcriptional memory lasts only 1.5 days after the last stimulation.

Table des Matières

Table des figures

Liste des tableaux

Liste des abréviations

AIC : Akaike Information Criterion

BIC : Bayesian Information Criterion

Db : Decibel

GO : Gene Ontology

Hz : Hertz

LASSO : Least Absolute Shrinkage and Selection Operator

MsL : Mechanosensitive Channel-Like

ODE : Ordinary Differential Equation

Pa : Pascal

QDR : Quantitive Disease Resistance

RNA : Ribonucleic Acid

RNA-Seq : RNA sequencing

SIMoNe : Statistical Inference for Modular Network

VAR : Vector autoregression

hpi : Hour post infection

1 Avant propos

Le fonctionnement des organismes vivants émerge d'une multitude d'interactions moléculaires, résultant en partie de l'expression génique. Pour mieux appréhender cette dynamique, la modélisation de la régulation de l'expression génique offre un potentiel considérable, permettant de synthétiser la complexité du vivant et d'extraire les paramètres essentiels. Les techniques actuelles de séquençage de l'ARN ouvrent la voie à la création de vastes ensembles de données décrivant l'évolution dynamique des paysages transcriptomiques, nécessitant des développements mathématiques pour concevoir des méthodes adaptées et performantes.

La modélisation de réseaux géniques a déjà démontré sa capacité à rendre intelligible la complexité de la biologie végétale. Par exemple, l'étude de Pomiès et al. (2017), utilisant un modèle de mélange, a permis de mieux comprendre la réponse du peuplier à des contraintes mécaniques répétées en analysant cinétiquement l'expression génique. Cette approche a confirmé le rôle central du facteur de transcription PtaZFP2, identifié de nouveaux gènes tels que PtaROPGEF14, et révélé que les premiers gènes modulés en réponse à des signaux mécaniques étaient liés à la résistance immunitaire.

Mon travail de thèse s'est principalement concentré sur l'inférence de réseaux de régulation dynamiques à l'échelle du génome, ainsi que sur sa mise en pratique pour explorer la mémoire transcriptionnelle associée à des stimulations sonores répétées.

Ma recherche se situe à la convergence des mathématiques appliquées et de la biologie végétale. j'ai travaillé sur la mise en place d'une méthode d'inférence de

réseaux de régulation dynamiques afin de modéliser la mémoire transcriptionnelle associée à des stimulations sonores répétées chez *Arabidopsis thaliana*. Cette modélisation de la mémoire transcriptionnelle a permis de prédire avec succès divers aspects du priming, par la suite confirmés par des expérimentations.

Les trois premiers chapitres de cette thèse présentent l'état de l'art relatif à la résistance quantitative (chapitre 2), aux données RNAseq (chapitre 3), et à l'inférence de réseaux (chapitre 4). Les chapitres suivants exposent les résultats obtenus au cours de cette thèse, commençant par la description des données d'expression liées aux répétitions de stimulations sonores chez des plantes saines et infectées (chapitre 5 et 6). Ensuite, nous abordons les méthodes de classification et d'inférence développées (chapitre 7) ainsi que leur évaluation sur des données simulées (chapitre 8). Enfin, le chapitre 9 (chapitre 9) présente les résultats de la modélisation de la mémoire transcriptionnelle.

2 La résistance quantitative: un contexte moléculaire cohérent pour tester le priming de l'immunité végétale par les ondes acoustiques.

Sommaire

Comme tous les êtres vivants, les plantes sont en contact permanent avec les micro-organismes de leur environnement. Au cours de l'évolution d'interactions physiologique et moléculaires complexes se sont développées. Ainsi, les interactions symbiotiques entre plantes et micro-organismes ont permis aux plantes de coloniser la terre il y a 450 millions d'années [Delaux and Schornack, 2021]. En parallèle, les interactions avec les pathogènes ont causé l'émergence d'un système de défense efficace et robuste. Cependant, en contexte agricole, les agents pathogènes des plantes constituent une menace majeure pour la sécurité alimentaire à l'échelle mondiale. La réduction des pertes de récolte dues aux attaques de pathogènes sous contrainte de limitation des traitements phytosanitaire, est une nécessité pour maintenir ou augmenter durablement les rendements. Le dérèglement climatique associé à l'augmentation de la population mondiale rendent primordiale la meilleure compréhension de l'immunité végétale.

Le contexte biologique de cette thèse est le priming de la réponse immunitaire. Le priming consiste à entraîner la plantes à répondre plus vite et plus fortement à des stress de toutes natures. L'entraînement de la plante se fait en exposant l'organisme à des stress non létaux pouvant êtres récurrents. Le priming implique donc la mémorisation des stresses passés. Bien que de nombreux exemples expérimentaux aient été rapporté, on ne connaît que peu de chose sur le fonctionnement de ces mécanismes de mémoire conduisant aux modifications des réponses végétales.

On pense souvent implicitement que le priming consiste à répondre mieux à une sollicitation de même nature que celles utilisées pour l'entraînement. Ainsi une plante exposée à des épisodes de sécheresse raisonnable répondra mieux à des épisodes de sécheresse sévère [Ling et al., 2018]. Ici nous nous intéresserons au cas du priming de la réponse immunitaire dit quantitative par des ondes acoustiques.

2.1 L'immunité végétale : un système moléculaire multi-couches

Le système immunitaire végétal est décrit comme plusieurs couches imbriquées. Chaque couche se différencie des autres par la nature des molécules associées au pathogène ou à son action.

2.1.1 La résistance qualitative

La forme de résistance la plus étudiée est la résistance dite qualitative qui groupe la PTI (Pathogen-Associated Molecular Pattern triggered Immunity) [Bacete et al., 2018] et l'ETI (Effector Triggered Immunity)[Roux et al., 2014]. Une partie de la résistance qualitative mobilise des récepteurs membranaires appelés PRR (Pattern Recognition Receptor). Les PRR ont généralement des domaines kinases et reconnaissent des motifs moléculaires conservés associés aux pathogènes appelés PAMP (Pathogen-Associated Molecular Patterns). Certains PRR peuvent également reconnaître des modèles moléculaires associés aux dommages crées par le pathogène (DAMP, Dammage Associated Molecular Patterns) le plus souvent dans la paroi. La reconnaissance des PAMP et des DAMP par les PRR déclenche une cascade moléculaire impliquant la génération de ROS (Reactive Oxygene Species); l'activation des MAP kinases (protéines kinases activées), la production d'hormones végétales comme l'acide salicylique, la modulation de gènes et le renforcement de la paroi cellulaire. On parle alors d'immunité déclenchée par PAMP ou PTI (PAMP Triggered Immunity) . Certains pathogènes peuvent infecter les plantes sans être reconnus par le PRR ou en supprimant le signal après l'infection. La suppression de cette PTI est souvent associée à la sécrétion de petites molécules qui facilitent l'infection par des agents pathogènes, appelées effecteurs de virulence.

La susceptibilité déclenchée par l'effecteur (ETS, Effector Triggered Susceptibility) est alors observée. Au cours de l'évolution, les plantes ont développé des récepteurs NLR intracellulaires (récepteurs de type NOD). Ce récepteur reconnaît spécifiquement les effecteurs produits par les pathogènes et est codé par des gènes de résistance appelés gènes R. La reconnaissance des effecteurs par les NLR peut être directe ou indirecte via l'engagement de cofacteurs [Cui et al., 2015], produisant des modulations conformationnelles qui activent les NLR. Lorsqu'ils sont activés, les NLR peuvent recruter directement ou indirectement des facteurs de transcription, permettant ainsi la reprogrammation transcriptionnelle des cellules et la mise en place de réponses protectrices appelées ETI (Effector-triggered immunity) [Cui et al., 2015]. Les voies de signalisation mobilisées par l'ETI sont complémentaires à celles déclenchées par la PTI et provoquent la mort cellulaire programmée. Cette

réponse dite hypersensible (HR) permet d'isoler les pathogènes des cellules saines de la plante. Cette résistance est dite qualitative car les plantes sont soient totalement résistantes aux maladies lorsqu'elles possèdent les protéines réceptrices adéquates soient totalement sensibles dans le cas contraire. [Jones and Dangl, 2006, Tao et al., 2003]. Cette forme de résistance est la plus étudiée pour des raisons biotechnologiques évidentes. Cependant, le recul sur l'introgression de gènes de résistance dans le génome de plantes sensibles questionne fortement l'efficacité de la démarche. En effet, il s'opère rapidement une sélection de pathogènes contournant le nouveau système de la plante. Le temps nécessaire à ce contournement est estimé à 3 ans. A titre comparatif, aujourd'hui, il faut une dizaine d'année pour créer et commercialiser une lignée génétiquement modifiée.

2.1.2 La résistance quantitative

En l'absence de résistance qualitative, c'est une autre forme de résistance qui est mobilisée pour faire face aux maladies. Cette forme de résistance qui mène à un phénotype de résistance incomplet est appelée résistance quantitative [Roux et al., 2014, Poland et al., 2009]. Elle se caractérise d'un point de vue phénotypique par une distribution continue de sensibilité à la maladie : les plantes sont plus ou moins résistantes. D'un point de vue moléculaire, cette forme de résistance est hautement multigénique et implique probablement la modulation de 20% du génome des plantes [Sucher et al., 2020] soit un nombre de gènes de l'ordre de grandeur de 10^4. Les gènes modulés par la QDR (quantitative disease resistance) ne sont pas spécifiques de l'immunité végétale et sont impliqués dans de nombreuses autres fonctions biologiques. Certains gènes NLR par la QDR, peuvent être exprimés et potentiellement induire la réponse hypersensible probablement au profit du pathogène [Barbacci et al., 2020]. C'est cette forme de résistance qui est mobilisée pour faire face aux attaques de champignons nécrotrophes comme *Botrytis cinerea* ou *Sclerotinia sclerotiorum* mais aussi face à certaines bactéries comme *Xanthomonas campestris* [Delplace et al., 2020].

Nous savons encore peu de chose sur les aspects spatio-temporels de la QDR et il a été montré récemment que l'implémentation spatiale de la QDR est en partie consécutive à la mécanoperception de signaux associés au pathogène. La lésion créée par le pathogène lorsqu'il a pénétré dans les tissus de l'hôte provoque le rééquilibrage des tensions mécaniques stockées dans les parois végétales sur un anneau entourant la lésion dont la largeur est du même ordre de grandeur que celui de la lésion. Le nouvel équilibre mécanique cause la réorganisation des microtubules corticaux dans les cellules seines à distance de l'infection. A cette auto-réorganisation spatiale est associée la modulation d'un grand nombre de gènes associés à la résistance qui participe à 40% du phénotype de résistance [Léger

et al., 2022]. La contribution de cette couche de résistance baptisée MTI (Mechano-signalling triggered Immunity) à la résistance globale peut également être modulée en appliquant des signaux mécaniques abiotiques à la plantes permettant d'obtenir des plantes plus résistantes ou plus sensibles aux maladies fongiques.

La sélection de variétés possédant une combinaison d'allèles de gènes de la QDR augmentant la résistance ou leur association avec des gènes R (pyramidage de gènes) semble être une solution permettant une plus grande durabilité de la résistance des variétés cultivées à divers pathogènes [Mundt, 2018, Pilet-Nayel et al., 2017]. En effet, des expériences ont montré que la résistance à *Leptosphaeria maculans* apportée par le gène Rml6 a une durabilité plus importante quand Rml6 est intégré à un fond génétique partiellement résistant (résistance quantitative) que quand il est intégré à un fond génétique sensible. La résistance apportée par Rml6 a été contournée en 3 ans dans le fond génétique sensible alors qu'elle a été contournée en 8 ans dans le fond génétique partiellement résistant [Brun et al., 2010, Delourme et al., 2014].

L'architecture moléculaire hautement mutligénique de la QDR contribue à rendre cette forme immunitaire tributaire des fluctuations de l'environnement. Les travaux menées dans l'équipe QIP (Quantitative Immunity in Plants) du LIPME (Laboratoire des Interaction Plante Microbes Environnemen) ont montré que la résistance d'*A. thaliana* au champignon nécrotrophe *S. sclerotiorum* diminue fortement lorsque les plantes sont cultivés dans un climat légèrement plus chaud que celui de leur niche écologique d'origine [Didelon, 2022].

Il a été avancé, sur la base d'analyses transcriptomiques, que les plantes perçoivent le son comme des signaux mécaniques [Ghosh et al., 2017]. Ce résultat est à prendre comme une proposition plutôt que comme une démonstration. En effet, les gènes de mécanoperception ne sont pas spécifiques à la mécanoperception et que l'on n'a jamais identifiés des sites mécano-perceptifs.

2.2 Effet du son sur les plantes

2.2.1 Une question initialement non biologique

La caractérisation de l'effet du son sur les plantes est une question qui est peu naturelle pour la biologie végétale. La rigueur avec laquelle une partie des études a été mené n'a pas aidé à piquer la curiosité des scientifiques. Pourtant cette idée à fait du chemin (Pour la science N°554, décembre 2023). Récemment, une série d'articles de qualité issues d'études biologiques sur la réponse des plantes aux ondes sonores a fait passer la question de la perception des ondes sonore du stade de curiosité mal placée à un élément potentiellement central à la co-évolution

plante/insectes et à l'adaptation de la plante à sa niche écologique [Appel and Cocroft, 2023b]. Même si ces perspectives sont exaltantes, la non connaissance des mécanismes perceptifs du son est un élément bloquant le développement de la thématique. Comme l'on a constaté que certain gènes de mécanoperception comme AtHSPRO2 (AT2G40000) ou AtTCH4 (AT5G57560) sont modulés suite à l'exposition d'une plante à une onde acoustique, il a été proposé que les ondes étaient mécano-perçues [Mishra et al., 2016]. Comme ces gènes ne sont pas spécifiques de la mécanoperception, la conclusion semble cavalière. Les études biomécaniques de l'interaction entre la structure végétale et les ondes acoustiques menées par le LIPME , MIAT (Mathématique et Informatique Appliquées de Toulouse) et le LAUM (Laboratoire d'Acoustique de l'Université du Mans) devraient permettre de répondre à cette question prochainement.

2.2.2 Rappel d'acoustique

Les principales caractéristiques des ondes sonores

Les ondes sonores sont des variations de pression qui se propagent dans les solides, les liquides ou les gaz **sans déplacement de matière**. Ces variations de pression sont décrites par :

Fréquence et longueur d'onde

la fréquence f qui représente le nombre des fois où la pression est maximale par unité de temps et est exprimé en hertz (Hz).

On peut exprimer des autres grandeurs temporelles en fonction de la fréquence :

$$T = \frac{1}{f}$$

où T est la période de la vibration qui a la dimension d'un temps et qui représente l'intervalle de temps qui sépare deux états de vibration identiques successifs. On peut également introduire :

$$w = 2.\pi f$$

la pulsation exprimé en rad/s.

La longueur d'onde représente la distance parcourue par l'onde pendant une période temporelle et a la dimension d'une distance (Figure : 2.1). Le longueur est inversement proportionnel à la fréquence : plus la longueur d'onde est grande, plus la fréquence est faible et inversement.

$$\lambda = \frac{c}{f}$$

avec c est la vitesse de l'onde dépendant du milieu [McCall, 2010] .

Le son audible perceptible par l'homme a des fréquences d'environ 20 Hz à 20 KHz, et au-dessus il s'agit d'ultrasons (Figure 2.2). Dans l'air à température et pression normales, les longueurs d'onde correspondantes des ondes sonores vont de 17 m à 17 mm.

La vitesse du son dans l'air est de 344m/s à 20 °C. À 1000 Hz (fréquence utilisée dans cette thèse) la longeur d'onde est de 34 cm. Les plantes d'*Arabidopsis thaliana*ayant une taille d'environ de 10 cm, on s'aperçoit que les fluctuations de pression autour de la plante sont faibles.

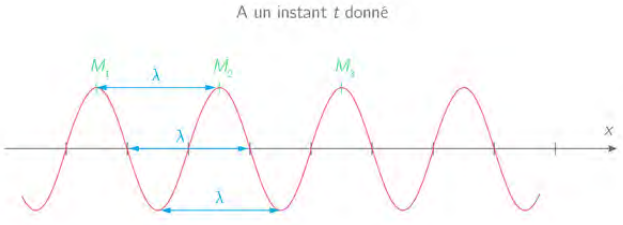


Figure 2.1: Définition de la longueur d'onde. source : www.kartable.fr/ressources/physique-chimie/cours/caracteristiques-des-ondes/22485

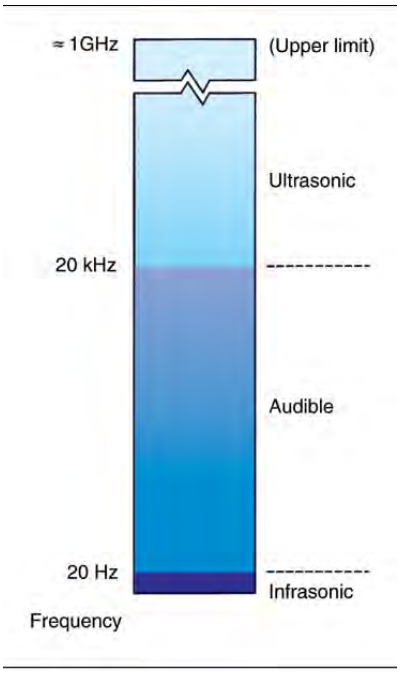


Figure 2.2: Échelle de fréquence [Shipman et al., 2012] : Le son audible perceptible par l'homme a des fréquences d'environ 20Hz à 20 KHz, pour des fréquences plus élevées on parle d'ultrason et d'infrason pour des fréquences plus basses.

vitesse de propagation d'une onde acoustique

Comme nous l'avons vu, le son se propage à une vitesse uniforme finie, indépendante de l'amplitude de l'onde et de sa forme. La vitesse s'exprime en fonction de la fréquence et de la longueur d'onde associée :

$$c = \lambda f$$

L'intensité acoustique

La propagation du son implique la propagation de l'énergie mécanique, **mais pas de la matière**. L'énergie transférée par unité de surface et par unité de temps est appelée intensité de l'onde sonore ou intensité acoustique.

L'intensité acoustique (ou puissance acoustique moyenne par unité de surface) correspond à l'amplitude de l'onde elle se calcule comme :

$$I = \frac{\rho A^2 \omega^2}{2} \lambda$$

Elle est de dimensions $[MT^3]$ et est un courant volumique d'énergie acoustique exprimée en watts par mètre carré (W/m2). Les grandes intensités acoustiques sont donc obtenues dans l'eau avec des ultrasons.

L'intensité sonore est souvent rapportée en décibel (dB). Le décibel n'est rien d'autre que la dixième partie du bel, unité créée vers 1923 par les ingénieurs des laboratoires Bell d'Alexander Bell. L'échelle logarithmique utilisée pour définir le décibel présente la particularité de bien décrire la perception humaine du bruit qui n'est pas linéaire (loi de Fechner). On peut exprimer l'intensité d'un son en décibels en calculant le rapport de l'intensité acoustique du son que l'on étudie à une intensité acoustique de référence fixée en utilisant la formule suivante:

$$L_I = 10 \log(\frac{I}{I_0})$$

Avec $I_0 = 10^{-12}\ W/m^2$ une intensité acoustique de référence fixée. Elle correspond au seuil d'audibilité par l'oreille humaine d'un son sinusoïdal de fréquence 1 kHz. Notons que le logarithme est décimal et que le seuil d'audibilité correspond à une surpression efficace qui vaut :

$$P_0 = \sqrt{I_0 \rho c} \sim 10^{-5} Pa$$

avec ρ la masse volumique de l'air. Une onde acoustique sinusoïdale d'intensité sonore 100 dB et de fréquence 1 kHz comme celle utilisée dans la suite de cette thèse correspond à une surpression de 1 Pa.

2.2.3 Le son : un signal mécanique pas comme les autres ?

Comme nous venons de le voir le son consiste en des fluctuations spatio-temporelles de faible pression à haute fréquence. Il est donc de nature mécanique mais est-il forcement mécano-perçu par la plante ? Les connaissances sur la mécanoperception des plantes ont rapidement évoluées ces 10 dernières années. On sait par exemple qu'à l'échelle macroscopique au moins ce sont les déformations mécaniques qui sont perçues comme l'indique le modèle S3M [Coutand and Moulia, 2000] :

$$S = \int\int\int_V k\,(\epsilon - \epsilon_0)\,dV$$

avec S la quantité de signaux perçus par la plante, k un paramètre décrivant la sensibilité perceptive de la plante, ϵ la déformation, ϵ_0 une déformation seuil en dessous de laquelle aucun signal n'est perçu et V le volume de cellule mécanoperceptive. On note qu'aucune valeur seuil n'est connue à ce jour et que des déformations de l'ordre de 10^{-4} sont perçues (Eric Badel, PIAF INRAe Clermont-Ferrand, communication personnelle). Cette forte sensibilité probable aux signaux mécaniques est en accord avec les résultats expérimentaux concernant la morphogénèse qui rapporte la réorganisation des microtubules associée à des variations de pression de turgescence faible [Sampathkumar et al., 2014]. L'ordre de grandeur des déformations créées par une onde acoustique à 100 dB est de l'ordre de 10^{-7} ce qui est inférieur à toutes les déformations minimales testées (10^{-4}). De plus la pression minimale à l'ouverture du canal mécanosensible (AtMSL10) est de l'ordre de grandeur de 1000 Pa [Peyronnet et al., 2014b] soit 1000 fois plus que la pression de l'onde acoustique ($\sim$ 1 Pa). Les connaissances actuelles ne garantissent donc pas que les plantes perçoivent le son.

De plus, le modèle S3M a été établi et démontré expérimentalement à de nombreuses reprises pour des signaux mécaniques quasi-statiques. Or, les ondes acoustiques sont des signaux mécaniques dynamiques à fréquence élevée. A l'heure actuelle, nous savons que le canal mécanosensible AtMSL10 favorise la perception de signaux mécaniques à une fréquence comprise en 0.3 et 3 Hz comme celles associées aux déformations crées par le vent [Tran et al., 2021] et n'avons aucune idée des bases moléculaires qui pourraient être impliquées dans la perception de signaux mécaniques à plus haute fréquence. Un modèle putatif impliquant les mécanismes connus de la mécanoperception a été proposé [Mishra et al., 2016]. Cependant, il n'est pas confirmé par aucune preuve expérimentale. Ce modèle implique entre autre la signalisation calcique. Pourtant, les étude faites au laboratoire ne détectent aucune variation de concentration en calcium (Xiong, Garcia, Barbacci unpublished data).

Ainsi, bien que le son soit un phénomène de nature mécanique il n'est pas du

tout certain qu'il soit mécano-perçu par la plante et on pourrait imaginer des perceptions alternatives :

- la perception de l'élévation de la température dans les tissus due à la fréquence élevée des ondes.
- la perturbation de la proprioception [Bastien et al., 2013, Moulia et al., 2021] par les faibles déplacements induits par l'onde incidente.

Des recherches sont encore à mener pour comprendre la perception du son par les plantes. La réponse à cette question permettrait de comprendre plus largement les connexions entre la mécanoperception et la réponse immunitaire. Par analogie avec la thigmomorphogénèse [Jaffe, 1973], qui décrit la régulation de la forme par les signaux mécaniques, nous pourrions parler de thigmoimmunité [Léger et al., 2022] pour évoquer la régulation de l'immunité par les signaux mécaniques.

Plusieurs travaux montrent déjà une connexion entre mécanoperception et réponse immunitaire. Des feuilles d'Arabidopsis carressées avec un pinceau ou roulées deviennent plus résistantes aux infections du champignon nécrotrophe *Botrytis cinerea* [Chehab et al., 2009, Benikhlef et al., 2013]. Des haricots (*Phaseolus vulgaris*) exposés au vent ont montré une augmentation de leur résistance à *Colletotrichum lindemuthianum* [Cipollini Jr, 1997]. La modélisation du réseau de régulation de la mécanoperception du peuplier *Populus tremula x P. alba* montre que les premiers gènes modulés sont des gènes impliqués dans la réponse immunitaire [Pomiès et al., 2017]. La PTI (Pathogen associated molcular pattern triggered immunity) pourtant basée sur l'interaction entre molécules spécifiques de plantes et de champignons est dépendante de la pression de turgescence et est régulée par les canaux mécanosensibles [Engelsdorf et al., 2018b].

2.2.4 Les plantes ne sont pas dures de la feuille

Des études ont montré que l'exposition des plantes à des ondes sonores répétées peut améliorer la résistance. Tout comme de nombreuses fonctions centrales de la plante. Par exemple, les répétitions de stimulations sonores favorisent la croissance des plantes en régulant les hormones de croissance des plantes. Les variations de concentration de l'acide indole-3-acétique (IAA) et de gibbérelline (Bochu et al.,2004; Ghosh et al.,2016), retardent la maturation des fruits de la tomate (Kim et al.,2015), augmentent l'expression des gènes liés à la photosynthèse (Jeong et al.,2008). Des travaux récents prouvent que les stimulations sonores activent la réponse immunitaire des plantes en modulant l'expression des gènes via les voies de l'acide salicylique (SA) et de l'acide jasmonique (JA) associés à la résistance aux

agents pathogènes (Ghosh et al.,2016). Ces effets sont résumés dans la figure 2.3 extraite de l'article de Jung et al,(Jung et al.,2018), et dans les tableaux 2.1 2.2.

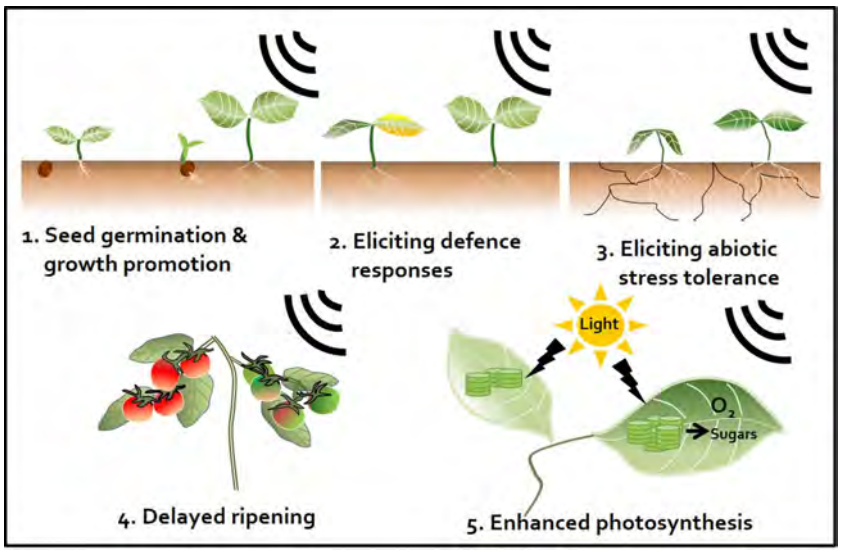


Figure 2.3: Effet du priming sonore sur les plantes [Jung et al., 2018].

plante	fréquence du son	effet du son	référence
Arabidopsis thaliana	500hz , 80db	active la réponse immunitaire des plantes en modulant l'expression des gènes dans les voies de l'acide salicylique et de l'acide jasmonique. Régulation de modificateurs de la paroi cellulaire	(Ghosh et al.,2016)
	200 Hz	augmentation de la production de Ca2+ cytosolique, de ROS et d'efflux de K+	(Rodrigo-Moreno et al.,2017)
	250 - 500 Hz	Activation des protéines liées à la réponse immunitare, la défense et la photosynthèse	Kwon et al.,2012
	1kh , 100db	defense contre l'infection à Botrytis cinerea via le voies de l'acide salicylique	(Choi et al.,2017a)
	500hz , 80db	defense contre l'infection à Botrytis cinerea via le voies de l'acide salicylique	Ghosh et al.,2017)
Tomate	1kh , 100db	retarde la maturation des tomates suite à la réduction de production de l'éthylène	(Kim et al.,2015, Kim et al.,2018)
Broccoli	250 hz - 1.5 khz	La teneur totale en flavonoïdes a augmenté de 35 % tout en améliorant l'efficacité antioxydante.	(Kim et al.,2020)

Table 2.1: Synthèse des effets des stimulations sonores répétées.

plante	fréquence du son	effet du son	référence
riz	400hz , 106db	augmentation du taux de germination, de la hauteur et des racines	(Bochu et al.,2003)
	250 hz - 1.5 khz	Tolérance à la séchresse.	(Jeong et al.,2014)
Fraise	0.1–1 kHz , 70–100 dB	Augmentation de la croissance et modification de la teneur en phytohormones.	(Hassanien and LI,2020)
	40 Hz-2 kHz	meilleure résistance aux maladies	(Qi et al.,2010)

Table 2.2: Synthèse des effets des stimulations sonores répétées.

2.2.5 Le priming

L'intégration des stimuli environnementaux détermine non seulement les réponses immédiates mais aussi futures des plantes aux environnements fluctuants. Les plantes apprennent du passé, se préparent pour les événements futurs. Les mécanismes biologiques par lesquels l'exposition aux signaux environnementaux rend les plantes plus résistantes aux événements futurs est appelés priming [Conrath et al., 2015]. Martinez-Medina et al. [2016] ont montré que les plantes primées sont plus résistantes aux stress environnementaux que les plantes non primées (ou naïves) car elles déclenchent plus rapidement des réponses de défense. Les mécanismes de priming sont donc des mécanismes centraux dans l'adaptation des plantes.

D'un point vue pratique, le priming de la défense des plantes est aussi une solution prometteuse pour la protection des cultures contre différents pathogènes. Le priming par des stimulations mécaniques pourrait être une des alternatives pertinentes aux produits chimiques [Ghosh et al., 2021] (Figure 2.4). L'utilisation de ce type de méthodes est néanmoins liée à une meilleure description et compréhension des conséquences biologiques. Par exemple, alors que les mécanismes d'acclimatation ont un effet majeur sur la thigmomorphogénèse [Martin et al., 2010c] ce mécanisme semble être étrangement absent des travaux concernant le priming par des signaux mécaniques. On note par ailleurs que nous n'avons pas observé d'acclimatation dans les études phénotypiques réalisées durant cette thèse (section.2.4).

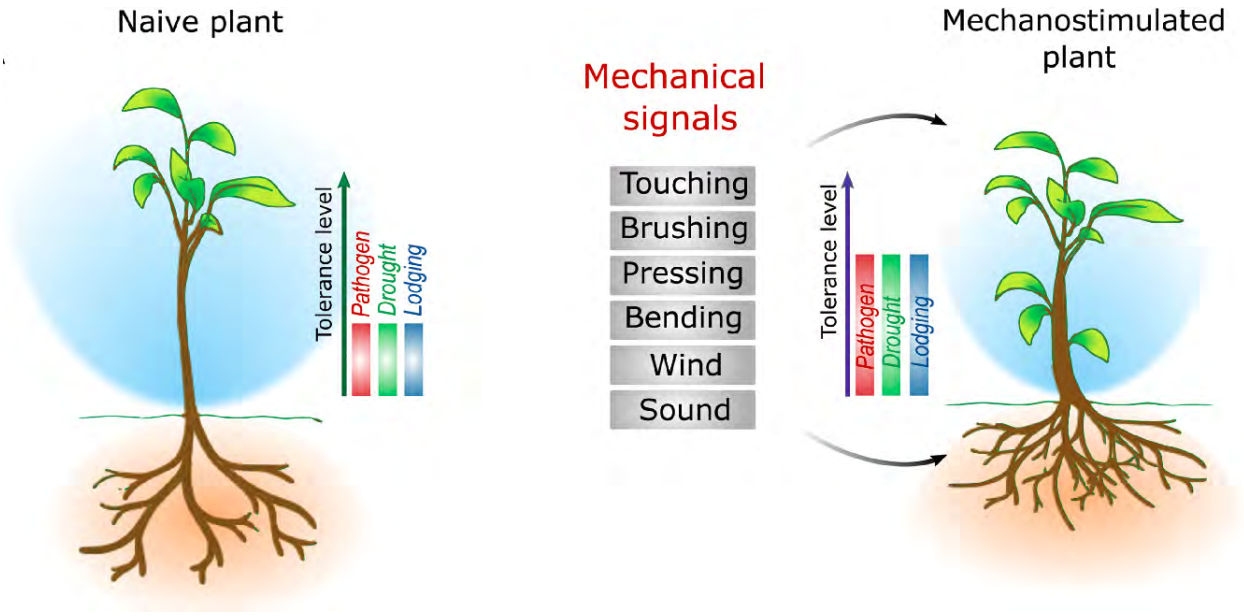


Figure 2.4: L'exposition répétée à des stimulations mécaniques modifie la morphologie des plantes, et les plantes mécano-stimulées présentent une plus grande tolérance au stress que les plantes non stimulées (source: [Ghosh et al., 2021]).

Le priming améliore globalement deux fonctions végétales : la résistance aux pathogènes et l'adaptation physiologique de la plante à son environnement.

Par exemple, Ling et al. [2018] montrent que les plantes d'*Arabidopsis thaliana* primées par des température élevées mais non léthales deviennent plus tolérantes à la chaleur. Cette tolérance à la chaleur est associée à une dé-répression du splicing au moment de la seconde exposition à la haute température. Au contraire, chez les plantes naïves non primées, les auteurs observent une forte mortalité associée à la répression du splicing. Des exemples de différentes méthodes de priming à différent stress sont présentés dans le tableau 2.3.

Le priming requiert l'existence d'une mémoire végétale. L'existence d'une mémoire chez les plantes a été parfois un sujet de controverse sans pour autant que de vraies raisons scientifiques remettent en cause son existence. Plus largement la mémoire est indissociable de l'irréversibilité des processus du vivant [Barbacci et al., 2015] et associée à la création d'entropie qui est une mesure physique de la flèche du temps. Par exemple l'arrêt de la croissance expansive (processus irréversible) par diminution de la pression de turgescence entraîne la mise en place d'une mémoire chimique composée de polysaccharides qui s'accumulent lors de la phase de basse pression. Lorsque la pression revient à un niveau normal, cette mémoire est utilisée pour compenser la perte de croissance [Barbacci et al., 2013].

Si les mémoires sont diverses, leur temps caractéristique l'est tout autant. L'ADN est sans doute la mémoire à plus long terme des organismes. Plus versatile, les

modulations épigénétiques sont à la base de mémoires allant de quelques minutes à plusieurs générations.
Les régulations épigénétiques interviennent à différent niveaux :

- Les histones qui forment avec l'ADN un complexe sujet à des modifications post-traductionnelles sur la partie N-terminale :
  - La méthylation : la présence d'un groupe méthyle associé des lysines correspond à un état répressif. C'est le cas de la tri-méthylation de la lysine 9 de l'histone H3 (H3K9). La méthylation d'autres résidus est associée à un état ouvert de la chromatine comme la méthylation de la lysine 4 de l'histone 3 (H3K4). Ces marques épigénétiques sont impliquées dans la thigmomorphogénèse du peuplier. Une flexion suffit à modifier le niveau global de H3K9 et les enrichissements en H3K9 et H3K4 des régions régulatrices des gènes sous-exprimés après la première flexion corrèlent avec le niveau d'expression de ces gènes [Ghosh et al., 2023b].
  - L'acétylation des lysines qui diminue l'interaction entre histone et ADN.
  - La phosphorylation , l'ubiquitination et la sumoyalation qui influent sur l'état d'ouverture de la chromatine.
- La méthylation de cystosine en 5-methylcytosine de l'ADN qui peut inhiber l'expression d'un gène.

Les modulations épigénétiques participent à la mémoire nécessaire au priming au même titre que d'autres acteurs de la régulation de l'expression des gènes comme les facteurs de transcription. Ces mécanismes de mémoire sont parfois appelés mémoire transcriptionelle [Avramova, 2019]. Dans cette thèse nous n'avons pas considéré l'épi-génome directement. Les modifications épigénétiques modulant l'expression des gènes, nous nous sommes servis de l'expression des gènes comme proxy aux régulations génétiques.

Méthode de priming	Plante	Stress effectué	Référence
Le vent	Haricot	champignon : *Colletotrichum lindemuthianum*	[Cipollini Jr, 1997]
La flexion	*A. thaliana*	champignon *B.cineria*	[Chehab et al., 2012]
Le Frottement	Fraise	champignon *B.cineria*	[Tomas-Grau et al., 2018]
Stress thermique	*A. thaliana*	bactérie : tomato DC3000	[Ling et al., 2018]
Stress salin	*A. thaliana*	bactérie : tomato DC3000	[Ling et al., 2018]
Vibration sonore	*A. thaliana*	champignon *B.cineria*	[Choi et al., 2017a]
Le Frottement	*A. thaliana*	champignon *B.cineria*	[Benikhlef et al.]

Table 2.3: Différentes méthodes de priming effectué sur des différentes plantes citées dans la littérature.

2.3 Effet des ondes acoustiques répétées sur la résistance d'*Arabidopsis thaliana*

Au cours de ces dernières années, plusieurs travaux des mêmes auteurs ont étudié l'effet de répétitions des ondes acoustiques sur la résistance d'*A. thaliana* . Ici, nous nous focalisons sur les 3 travaux qui se sont intéressés à l'effet du son sur l'expression des gènes :

- Ghosh et al. [2016] s'intéressent à l'effet de la fréquence sur le transcriptome chez la plante saine.
- Choi et al. [2017a] s'intéressent à la QDR de plantes soumises au son pendant 10 jours.
- Ghosh et al. [2017] s'intéressent à l'expression des gènes en fonction de temps d'exposition court chez la plante saine.

2.3.1 Effet de la fréquence sur l'expression des gènes

Le but de cette étude de Ghosh et al. [2016] était d'identifier l'effet immédiat d'onde sonore sur le transcriptome d'*A. thaliana* en fonction de la fréquence. Des plantes de 20 jours ont été exposées à des ondes sonores de différentes fréquences (250, 500, 1000, 2000 et 3000 Hz) en maintenant l'intensité sonore constante (80 dB) pendant 1 heure. Les transcriptomes ont été analysés par microarray à la fin de l'exposition.

Le plus grand nombre des gènes différentiellement exprimés était associé aux ondes de fréquence 500 Hz. Une analyse d'ontologies a été effectuée sur tous les gènes différentiellement exprimés pour les 5 fréquences. La majorité des gènes s'est révélée impliquée dans les réponses aux stress et dans les voies de transduction du signal. 59 gènes ont été modulés dans au moins trois des fréquences utilisées. Ces gènes ont été classés dans des différents groupes selon leurs fonctions biologiques (2.5) :

- Groupe A. regroupe 23 gènes répondant aux signaux mécaniques [Lee et al., 2005] sur-exprimés par la stimulation sonore.
- Groupe B. Les gènes liés à la signalisation .
- Groupe C. Les facteurs de transcription.
- Groupe D. Les gènes participant à l'homéostasie rédox.
- Groupe E. 10 gènes de biosynthèse.
- Groupe F. Gènes associés à la défense. On note qu'une partie des gènes de défense est aussi activée par des stimulations mécaniques. 3 voies sont modulées l'acide jasmonique, l'acide salicylique et la réponse à la chitine, polysaccharide composant majoritairement les parois des hyphes fongiques.
- Groupe G. Gènes impliqués dans d'autres processus.
- Groupe H. Gènes inconnus.

Ces résultats montrent pour la première fois que les variations phénotypiques observées dans d'autres travaux résultent des changements de programmation transcriptomique de grande ampleur induit par les stimulations sonores. La corrélation entre fréquence de l'onde stimulatrice et le nombre de gènes exprimés reste inexpliquée bien que d'importance pour les recherches futures sur la perception du son par les plantes.

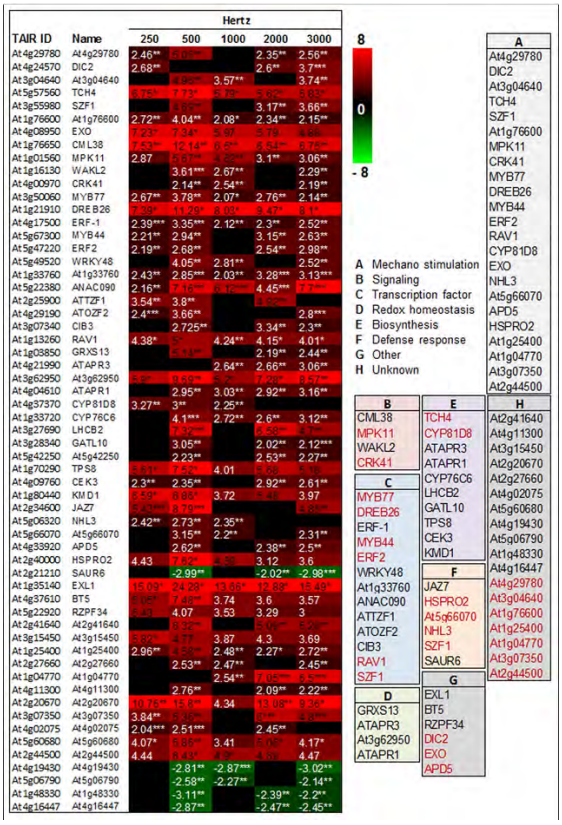


Figure 2.5: Heatmap des 59 gènes différentiellement exprimés par au moins trois fréquences sonores distinctes. [Ghosh et al., 2016]

2.3.2 Relation entre niveaux d'expression et temps de stimulation

Dans l'article de Ghosh et al. [2017] qui a pour objectif principal d'identifier les bases moléculaires de la perception du son par les plantes, des plantes d'*A. thaliana* âgées de 20 jours ont été exposées à des stimulations sonores de fréquence constante 500Hz, d'intensité 80 dB durant différentes périodes temporelles : 10, 30 minutes, 1h et 2 h. Le but de cet étude était d'étudier le comportement d'expression précoce des gènes d'*A. thaliana* modulés par le son [Ghosh et al., 2017]. L'ensemble de ces 17 gènes est divisé sans surprise en deux groupes :

- le premier groupe contient les gènes qui sont sur-exprimés après 1h ou 2h (CML38, TPS8, BT5, RZPF34, LHCB2, At3g07350, At2g44500 et At2g20670).
- le deuxième groupe contient les gènes sur-exprimés à la fois après 1h et 2h (CYP76C6,MPK11, MYB77, RAV1, KMD1, DREB26, HSPRO2, At1g76600 et At1g25400).

Les auteurs s'intéressent ensuite à 13 gènes putativement mechano-sensibles

extraits de la liste donnée en 2005 par Lee et al. [2005]. On compare alors l'expression des gènes des plantes soumises au son à celle des plantes caressées avec un pinceau par Lee et al. [2005]. Seulement 2 gènes parmi les 17 gènes (MSL3 et MSL7) étaient sur-exprimés dans les deux cas. Un seul gène (MCA2) présentait un modèle d'expression opposé.

La conclusion des auteurs est que la perception du son est sans doute différente de la mécanoperception. Cette conclusion est à prendre avec beaucoup de précautions puisqu'elle comporte de nombreuses zones floues. Les limites sont d'ordre

- biologiques : l'âge des plantes n'est pas le même dans l'étude de Lee [Lee et al., 2005] et dans l'étude de de Ghosh [Ghosh et al., 2017].
- métrologiques : comment faire une stimulation normalisée avec un pinceau [I], pour combler le manque de puissance statistique associé à l'expression de 17 gènes mesurée par microarray, l'expression des gènes est réanalysée par q-RT PCR, methode connue pour la grande variance de ses résultats
- conceptuelles : les gènes de la mécanoperception ne sont pas spécifiques à la mécanoperception (y compris les gènes TCH - touch-) comme le montre d'ailleurs l'article de Lee et al. [Lee et al., 2005]

Le principal enseignement de cette étude est que la compréhension de la perception du son par les plantes doit intégrer la physique de l'interaction entre le son et la plante puisque les approches de biologies moléculaires seules s'avèrent trop limitées.

2.3.3 Effet des ondes sonores sur la réponse d'*A. thaliana* aux champignon nécrotrophe *Botrytis cinera*

Pour étudier l'effet des ondes sonores sur la résistance d'*A. thaliana* aux champignons nécrotrophes, les plantes âgées de 14 jours ont été exposées au son 10 jours pendant 3h par jour. La fréquence de 1000 Hz a été maintenue constante à une intensité de 100 dB [Choi et al., 2017a]. Suite au traitement sonore, les plantes ont été infectées par *Botrytis cinerea*. L'analyse des transcripts s'est faite par microarray à la fin de la stimulation avant l'infection et 12h et 24h après inoculation par le champignon (hpi).

L'analyse phénotypique montre un gain important de résistance. 72 heures après l'infection (hpi), les plantes exposées au son ont des lésions presque 2 fois plus petites que les plantes contrôles. Ici le critère phénotypique considéré est une variable intégrative du processus infectieux : la taille de lésion mesurée dépend du

[I] on peut trouver quelque tentative pourtant

temps de latence ainsi que de la vitesse de la nécrose. Dans notre étude (section.2.4) nous avons utilisé un autre critère qui est la vitesse de croissance de la nécrose. Notre critère phénotypique ne tient pas compte du temps de latence que nous avons trouvé trop fortement associé à la capacité du champignon à former des structures de colonisation comme des appresoria ou des "coussins d'infection" et peut descriptif de la réponse immunitaire de la plante [Barbacci et al., 2020].

Le plus grand nombre des gènes différentiellement exprimés est détecté 24h après l'infection. L'analyse d'enrichissement des processus biologiques de ces gènes indique qu'un plus grand nombre de gènes liés à la défense étaient sur-exprimés 12 hpi (heures après l'infection) dans les plantes traitées par rapport aux plantes contrôles. Cet enrichissement en ontologie de défense était déjà présent chez les plantes exposées mais non infectées. A 24h, les gènes liés à la réponse de défense induisent la résistance systémique acquise (systemic acquired resistance) et à la réorganisation de la paroi cellulaire. Cette étude d'enrichissement montre aussi que plusieurs gènes de réponse aux stress abiotiques (osmotique, sel, intensité lumineuse élevée et chaleur) étaient sous-exprimés dans les plantes exposées au son 24 hpi. Pour analyser les gènes qui évoluent de la même façon au cours du temps, une classification hiérarchique est appliquée sur les 280 gènes différentiellement exprimés. 7 classes ont été identifiées sur la base de la cinétique d'expression de ces gènes. Une étude d'enrichissement de chaque classe a été réalisée. Les gènes des classes 3 et 4, sur-exprimés 12 hpi, étaient essentiellement des gènes de défense. Les gènes liés à la résistance systémique acquise et à l'organisation de la paroi cellulaire appartiennent aux classes 5 et 6 où les gènes étaient sur exprimés à 24 hpi. Les résultats de la classification et de l'analyse d'enrichissement sont présentés dans la figure 2.6.

Les analyses transcriptomiques de ces données montrent que l'acide jasmonique (JA) qui joue un rôle essentiel dans l'immunité des plantes face aux champignons nécrotrophes, est antagoniste de la voie de défense activée par l'acide salycilique (SA) impliqués dans la défense aux bactéries. L'analyse des gènes impliqués dans l'activation des hormones de défense JA et SA montre que le niveau de SA des plantes exposées au son a augmenté lors de l'infection (entre 12hpi et 72hpi). De plus, une concentration accrue de SA a été observée après le traitement par SV par rapport au contrôle non traité à 0 h suggérant que la variation hormonale est liée au son et pas à la maladie. Le niveau de JA dans les plantes exposées au son est resté stable pendant l'infection, alors qu'il a été progressivement induit dans les plantes contrôles (Figure.2.7).

Ces résultats montrent pour la première fois que le gain de résistance phénotypique conféré par l'exposition des plantes à des stimulations sonores répétées est associé à la reprogrammation transcriptomique de deux voies hormonales majeures

de la défense.

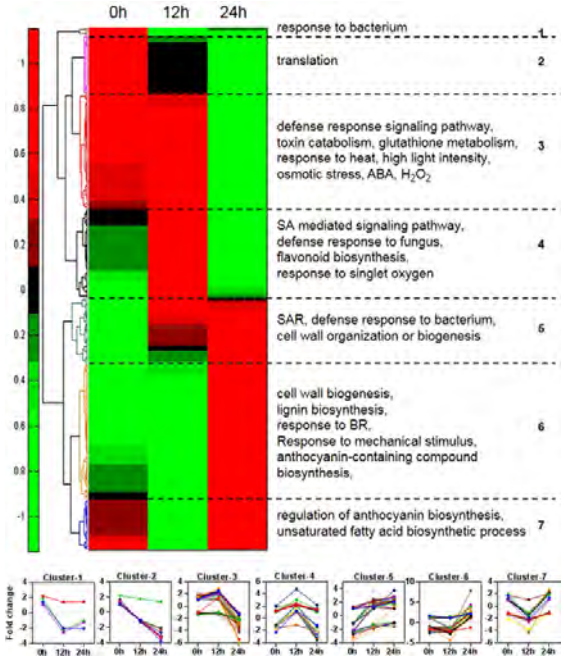


Figure 2.6: Classification hiérarchique de gènes différentiellement exprimés après l'infection par *B. cinerea* chez *Arabidopsis thaliana* exposée au son avec une fréquence de 1 kHz. L'analyse d'enrichissement des ontologies de processus biologiques a été mentionnée pour chaque groupe. [Choi et al., 2017a].

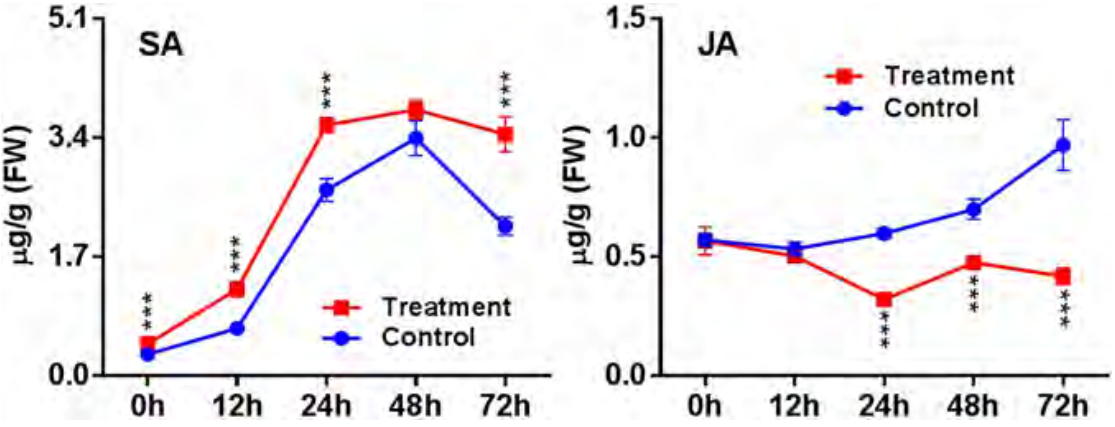


Figure 2.7: Contenu en acide salicylique (SA) et en acide jasmonique (JA) après l'infection par *Botrytis cinerea* chez *Arabidopsis thaliana* exposée au son avec une fréquence de 1 kHz [Choi et al., 2017a].

2.4 Effet phénotypique des ondes acoustiques répétées sur la résistance d'*Arabidopsis thaliana* au champignon nécrotrophe *Sclerotinia sclerotiorum*.

Le priming par répétition de stimulations sonores a été établi phénotypiquement et on a étudié les modulations transcriptionnelles des plantes primées pendant l'infection après le priming [Ghosh et al., 2017]. La question de comment se construit le priming avec les répétitions de stimulations sonores reste entière.

Pour comprendre l'effet des stimuli acoustiques répétés sur la défense des plantes nous avons évalué quantitativement l'effet phénotypique d'ondes acoustiques répétées sur la résistance d'*Arabidopsis* au champignon nécrotrophe *Sclerotinia sclerotiorum*. Des plantes *A. thaliana* âgées de quatre semaines ont été exposées pendant 1 à 10 jours à stimulations acoustiques d'amplitude 100 dB (1Pa) et 1 KHz pendant 3h par jour (synchronisées sur le début du jour). Les plantes ont été cultivées en jour court (9h) à température constante (23°C) et avec une humidité de 70%. La quatrième semaine, elles ont été disposées en phytotron dans des enceintes transparentes munies d'un haut parleur relié à un amplificateur de puissance et à un ordinateur coordonnant le début et l'arrêt de la période de sollicitation.

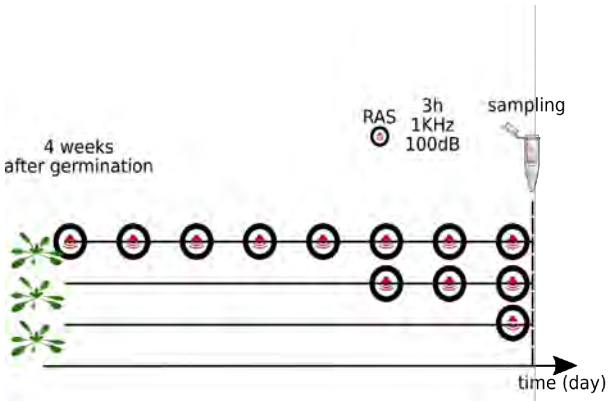


Figure 2.8: Protocole d'échantillonnage expérimental : les plantes ont été exposées à 0 à 8 stimuli acoustiques (1 KHz, 100 dB, 3h/jour). Des échantillons de feuilles ont été extraits en même temps pour toutes les plantes.

2.4 Effet phénotypique des ondes acoustiques répétées sur la résistance d'*Arabidopsis thaliana* au champignon nécrotrophe *Sclerotinia sclerotiorum*.

À la fin des sollicitations, on évalue l'effet de la répétition de stimulations acoustiques en infectant avec *S. sclerotiorum* toutes les feuilles de largeur supérieure à 5 mm de chaque plante. Les feuilles sont détachées manuellement et disposées dans un dispositif de phénotypage appelé Navautron permettant de maintenir un haut degré d'humidité [Barbacci et al., 2020]. Ce dispositif permet de prendre des images des infections à intervalle régulier (ici 10 min). Un code python permet d'extraire la progression de la maladie (chlorose) pour chaque feuille. La courbe de développement de la maladie est une courbe en "S" [II]. La figure 2.9 décrit l'analyse de la résistance quantitative aux maladies (QDR) à *Sclerotinia sclerotiorum* avec le système Navautron.

[II] mais pas une vraie sigmoïde au sens mathématique

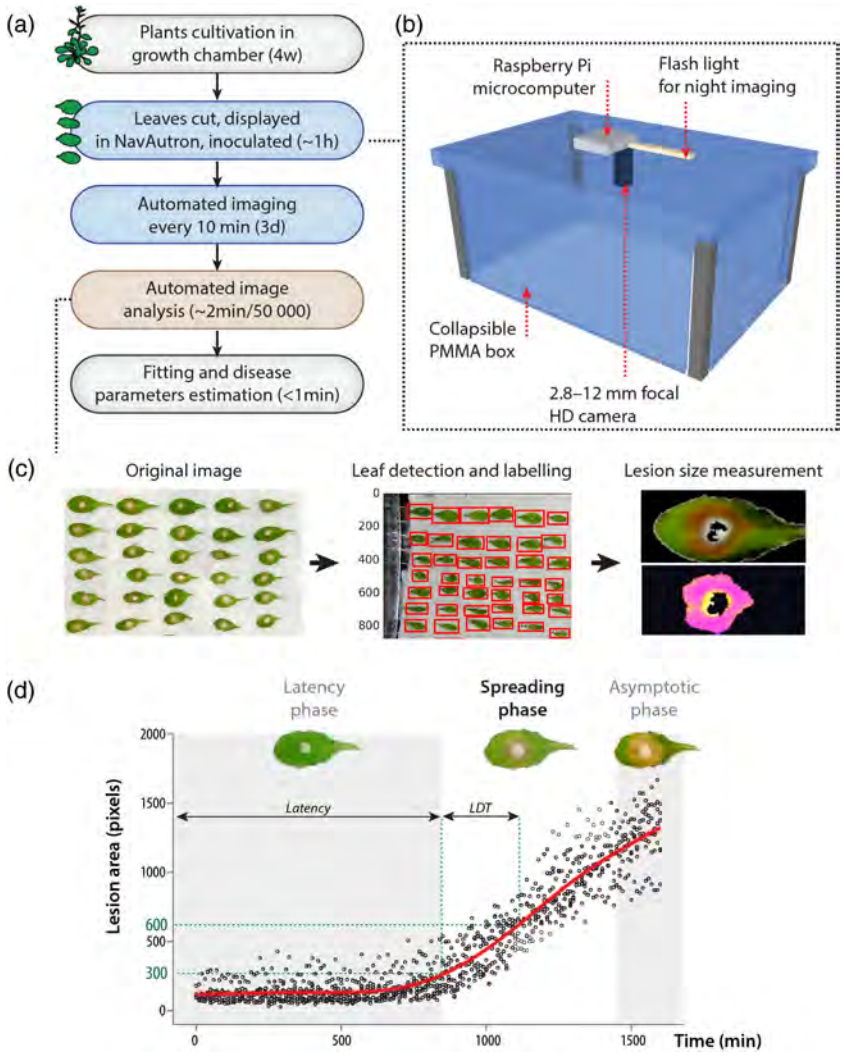


Figure 2.9: Analyse de la résistance quantitative aux maladies (QDR) contre *Sclerotinia sclerotiorum* avec le système de phénotypage Navautron.
(a) Pipeline de phénotypage. (b) La configuration Navautron : chaque Navautron est constitué d'un boîtier en plastique transparent équipé d'un micro-ordinateur, d'une caméra haute définition et d'un flash LED. (c) Les trois étapes principales de l'analyse d'image automatisée. (d) Cinétique typique du développement des lésions de la maladie de *S. sclerotiorum* sur *A. thaliana* , illustrant la phase de latence, la phase de propagation et la phase asymptotique [Barbacci et al., 2020].

L'infection se décompose en 3 phases : le temps de latence qui correspond au temps nécessaire à l'apparition des premiers symptômes et le taux de croissance de la phase exponentielle même si seulement deux des trois phases de la progression sont pertinentes pour décrire le processus infectieux (la phase de latence décrivant la taille de la feuille). L'étude menée sur plusieurs génotypes d'*A. thaliana* et de *S. sclerotiorum* indique que le taux de croissance exponentielle est la variable la

plus pertinente pour décrire le niveau de résistance de la plante [Barbacci et al., 2020]. Plus le taux est faible et plus la plante est résistante. Inversement, un taux élevé indique une plante sensible à l'infection. De plus, la détermination du temps de latence est approximative puisque fortement dépendant de la méthode utilisée pour détecter la nécrose qui se fait par analyse colorimétrique (un pixel fait partie de la nécrose si sa composante rouge est supérieure à sa composante verte). Pour quantifier la résistance de la plante, il est nécessaire de trouver une mesure non-dépendante du zoom de la caméra ou de la taille des feuilles. Deux critères sont possibles : un temps caractéristique de doublement de taille de lésion [Barbacci et al., 2020] ou directement le taux de croissance de la phase exponentielle. Dans cette thèse, nous avons choisi d'utiliser le taux de croissance qui est obtenu par ajustement de la droite liant taille de lésion et temps dans l'espace temps-log taille de lésion à l'aide du package "segmented" de R [Muggeo et al., 2008].

Ainsi, plus le taux de croissance est élevé et plus la plante est sensible[III] ou moins résistante à l'infection. Pour des raisons communautaires, on utilise préférentiellement la sensibilité.

Dans un premier temps, nous nous intéressons au rôle des trichomes dans l'effet de la répétition de signaux mécaniques sur un temps fixé à 10 jours [Ghosh et al., 2016]. Les trichomes ont été proposés comme organes déterminant dans la perception du son par les plantes [Liu and Jiao, 2017]. Des travaux plus récents proposent qu'ils sont également impliqués dans la perception de la pluie [Matsumura et al., 2022]. En effet, leur forme élancée les rendrait particulièrement sensibles aux flexions mécaniques induisant des fluctuations de tension dans les parois potentiellement perçues par les cellules végétales [Hamant and Haswell, 2017]. Dans le cas du son [Liu and Jiao, 2017], il n'existe aucune preuve expérimentale et cette proposition est faite sur la base d'un modèle biophysique très fortement dépendant de la paramétrisation de l'épaisseur de la paroi dans les trichomes. Or ce modèle semble faux (cf. supplementary de Liu and Jiao [2017]). Pour tester si la présence des trichomes a un effet sur l'effet de 10 jours de stimulations acoustiques répétées, nous avons utilisé 2 lignées mutantes dans le fond génétique de Col-0. arpc-4 est une lignée mutante affectée dans la polymérisation de l'actine [Badet et al., 2019]. Il en résulte que les trichomes de ce génotype exhibent une forme altérée moins sensible à la flexion. gl-1 est une lignée mutante glabre exhibant peu ou pas de trichomes [Hauser et al., 2001]. Les résultats que nous obtenons montrent que pour chaque génotype les stimulations acoustiques provoquent une diminution substantielle de la sensibilité à l'infection par le champignon (Figure 2.10 .A). Ainsi, les trichomes apparaissent très secondaires pour expliquer l'effet de stimulations

[III] susceptibility sur les figures.

acoustiques répétées sur la modulation de la réponse immunitaire.

Nous avons ensuite testé l'effet du nombre de répétitions sur la modulation de la résistance en utilisant la lignée sauvage Col-0. Les mesures de phénotype ont été faites après 1, 2, 3 et 10 jours de stimulations acoustiques (Figure 2.10 .B). Après 1 jour de son, on observe une faible augmentation de la sensibilité à l'infection puis une diminution qui devient importante après 3 jours et reste stable jusqu'à 10 jours. Ainsi, après 3 jours, le gain de résistance est d'environ 25% ce qui est considérable en comparaison des quelques % observés pour les gènes de QDR connus. Il est a noter que ces résultats suggèrent que trois jours sont nécessaires pour sensibiliser ou potentialiser la plante et qu'aucune phase d'acclimatation ou de désensibilisation n'est observée jusqu'à 10 jours. Ces résultats contrastent donc avec ce qui est connu concernant la mécanoperception des signaux mécaniques crées par flexions répétitives qui entraînent une acclimatation rapide de la plante [Martin et al., 2010a].

Ces résultats suggèrent que la diminution de la sensibilité induite par le son est un processus dynamique nécessitant l'intégration temporelle de la réponse de la plante à la perception du son indépendamment des trichomes. Pour comprendre les bases moléculaires associées à ces changements phénotypiques, nous avons généré un jeu de données de RNAseq décrivant l'évolution de l'expression génétique en fonction du nombre de stimulation acoustiques.

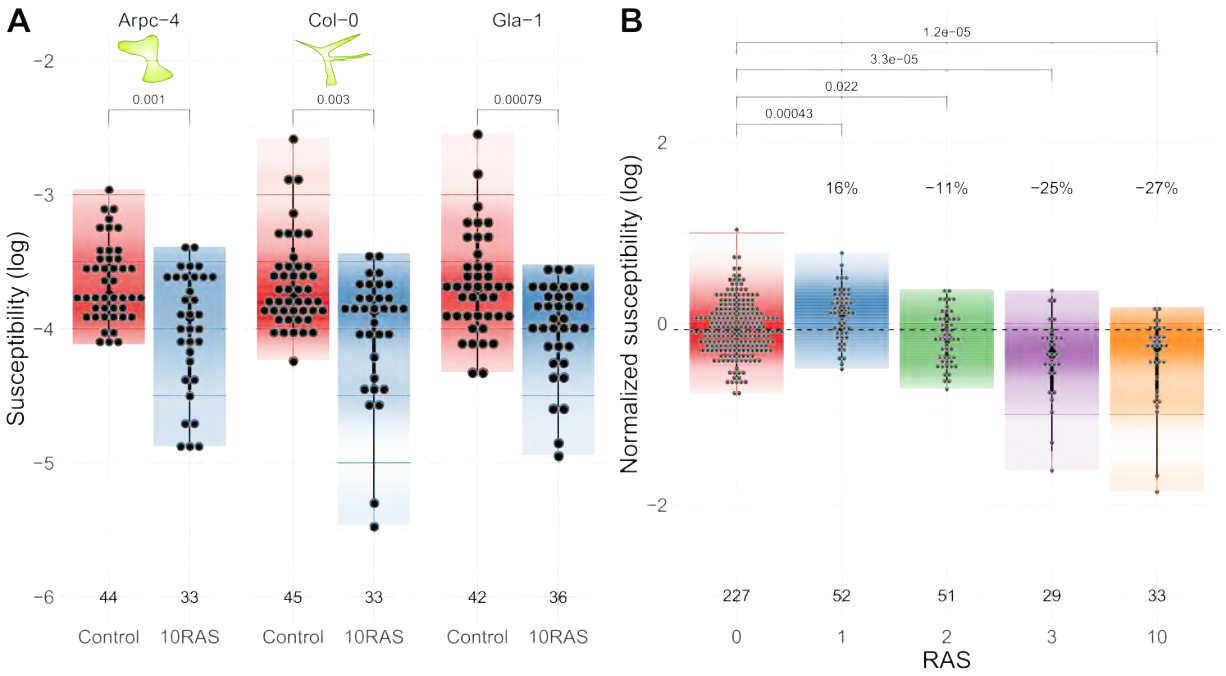


Figure 2.10: Effet de répétition de stimulations acoustiques sur la sensibilité des plantes *A. thaliana* à *S. sclerotiorum*. **A** Une perte de sensibilité de 25 % a été associée à des stimulations de 3 h/jour pendant 10 jours indépendamment des génotypes affectés et des trichomes. **B** Des comparaisons de la sensibilité moyenne ont été effectuées avec des plantes non stimulées. La perte de susceptibilité associée aux répétitions de stimulations acoustiques est conférée après 3 stimuli et reste stable.

3 De l'ADN à l'ARN et son analyse

Sommaire

3.1 Expression géniques

3.1.1 De l'ADN à l'ARN

Les cellules sont des unités biologiques fonctionnelles qui composent tous les organismes vivants. Ce sont les plus petites unités de vie capable de se reproduire. L'activité cellulaire est coordonnée et programmée sur la base des informations stockées sur une grande molécule appelée acide désoxyribonucléique (ADN). L'ADN est présent dans toutes les cellules vivantes et contient toutes les informations dont un organisme a besoin pour se développer et fonctionner. Il porte l'information génétique (génotype) qui constitue le génome d'un organisme. La molécule d'ADN est une double-hélice droite qui est composée de deux brins complémentaires (Figure 3.1). Chaque brin d'ADN est constitué de quatre types de nucléotides : adénine (A), cytosine (C), guanine (G) ou thymine (T). Les nucléotides trouvés dans un brin ont des nucléotides complémentaires dans l'autre brin avec lesquels ils peuvent interagir (A complète T, G complète C).

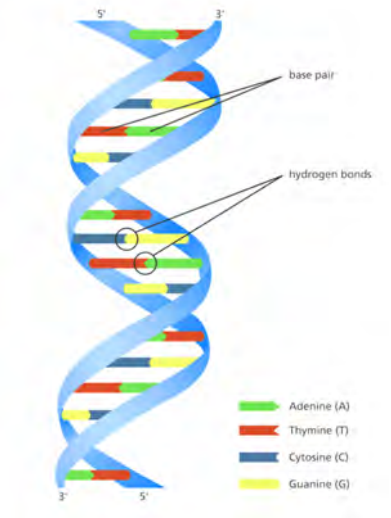


Figure 3.1: Schéma de l'ADN. Source : Genome Research Limited.

Un gène est une unité d'information génétique codée sous la forme d'une séquence de nucléotides et correspond donc à une petite portion du génome. Chez les eucariotes, les gènes sont constitués d'une alternance d'exons (régions codantes) et d'introns (régions non-codantes). Le nombre de gènes varie d'un organisme à l'autre, quelle que soit la taille du génome, allant de centaines à des dizaines de milliers de bases.

La première étape de la synthèse protéique est la transcription de la séquence codante en ARN. Comme l'ADN, l'ARN est un transporteur moléculaire d'informations génétiques. L'ARN est constitué d'un seul brin. Les informations sont également codées à l'aide de quatre nucléobases. Ici, la thymine (T) est remplacée par l'uracile (U). Il existe de nombreux types d'ARN, classés selon leur fonction. Certains ARN contiennent des informations génétiques qui codent pour des protéines. Ces ARN, dits codants, sont appelés ARN messagers (ARNm). L'étape de construction de protéines à partir d'ARNm est appelée traduction [Crick, 1970]. Néanmoins, d'autres ARN jouent des rôles importants dans la fonction cellulaire et sont particulièrement impliqués dans la régulation de l'information et de l'activité cellulaire. Les étapes de transcription et de traduction illustrées dans la figure 3.2 sont régulées par d'autres protéines en fonction de l'état cellulaire. C'est par exemple le cas des facteurs de transcription qui contrôle le taux de transcription de l'information génétique de l'ADN à l'ARN messager, en se liant à une séquence d'ADN [Latchman, 1997]. Le rôle principal des facteurs de transcription est d'activer et de désactiver les gènes afin qu'ils soient exprimés au bon moment et en quantité suffisante dans les cellules d'intérêt tout au long de la vie de la cellule et de l'organisme. Pour les cellules, les populations d'ARNm ou de protéines sont caractéristiques de l'état à un moment donné. Il est donc possible de mesurer l'abondance de tous les ARNm ou de toutes les protéines afin de mieux comprendre ce qui se passe à l'intérieur de la cellule.

Un transcriptome est un ensemble d'ARNm produits par un tissu ou une cellule à un moment précis et dans des conditions précises. Un transcriptome peut être considéré comme un miroir de toutes les protéines produites par une cellule. La quantification du transcriptome dans des conditions spécifiques permet potentiellement d'identifier les gènes transcrits, déterminer les mécanismes de régulation de l'expression de ces gènes et par la suite identifier les réseaux de régulation.

Un protéome correspond à toutes les protéines fabriquées à partir d'ARNm. Les protéines régulent la synthèse, la destruction et l'activité des métabolites et varient potentiellement d'une cellule. Il n'y a pas de corrélation directe entre les niveaux d'expression d'ARNm et les niveaux d'expression de protéines dérivées. Par exemple, certains processus de régulation peuvent affecter le taux de formation et de dégradation des protéines pendant la traduction. De plus, il est courant que des protéines soient exportées en dehors de leurs cellules productrices. Cependant, la quantité des protéines est très difficile à mesurer, car les protéines sont instables et se dégradent rapidement dans les cellules. L'association d'une protéine particulière à un emplacement dans le génome est plus difficile. Mesurer la quantité d'ARN mature et stable est plus facile. Par conséquent, cette quantité est un bon proxy

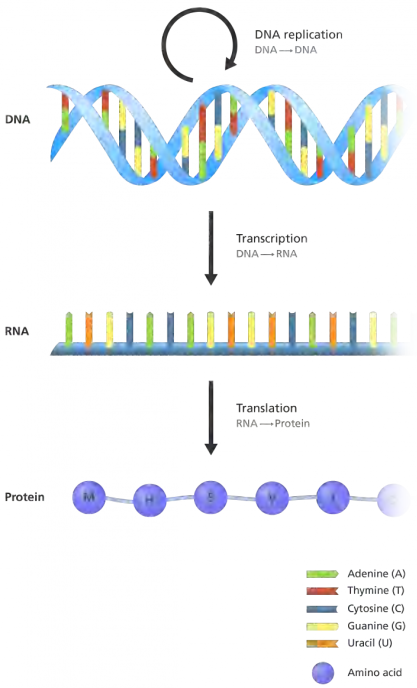


Figure 3.2: Dogme central de la biologie moléculaire. Les gènes sont exprimés dans les cellules par transcription de l'ADN en ARN messager, qui est ensuite traduit en protéine. Source : Genome Research Limited.

pour comprendre l'activité cellulaire.

3.1.2 Régulation de l'expression des gènes

Avec les facteurs de transcription, l'épigénétique est un régulateur de l'expression des gènes. Deux organismes contenant exactement le même génome, peuvent exprimer différemment les mêmes gènes et donc avoir des caractéristiques différentes. Les changements épigénétiques régulent l'expression des gènes, et donc l'activité des organismes.

Les mécanismes épigénétiques consistent en des modifications directes de l'ADN, soit par méthylation (ajout d'un groupement méthyle à l'un des nucléotides qui composent l'ADN), soit des modifications de protéines de liaison à l'ADN appelées histones. Ces changements épigénétiques affectent la liaison des facteurs de transcription et de l'ARN polymérase, affectant ainsi indirectement l'expression des gènes. Il existe une relation complexe entre les facteurs de transcription et la machinerie épigénétique. Par exemple, les facteurs de transcription peuvent recruter des enzymes capables de supprimer ou d'ajouter des marques épigénétiques.

Il existe trois types de marques épigénétiques : la méthylation de l'ADN, les modifications des histones et les interactions avec l'ARN non-codant. Tous ces marqueurs sont stables et ont la propriété de se transmettre lors de la division cellulaire même s'ils sont modifiables et/ou réversibles. Ces changements peuvent se produire spontanément ou en réponse à l'environnement. Les marques épigénétiques sont des mécanismes de mémoire d'événements passés qui peuvent expliquer par exemple les différences observées entre clones partageant la même séquence d'ADN. Grâce aux nouvelles technologies de séquençage, il est désormais possible de lire l'intégralité du code épigénétique des plantes.

L'étude des mécanismes épigénétiques est importante pour mieux comprendre le priming puisqu'ils peuvent se transmettre d'une génération à l'autre permettant à la prochaine génération d'être mieux préparée au stress. C'est le cas par exemple de stress associé à la lumière, la température ou la qualité du sol [Youngson and Whitelaw, 2008]. Dans quelques cas décrits dans la littérature, l'effet est perdu après deux générations.

3.1.3 Quantification du nombre de transcrits

L'ARN est extrait des tissus ou des cellules et transcrit en ADN complémentaire (ADNc). Une étape importante est la réplication en masse de séquences d'ADN pour produire des quantités suffisantes pour quantifier les transcrits. À cet effet, on utilise la réaction en chaîne par polymérase (PCR) [Mullis and Faloona, 1987]. Il est alors possible de quantifier la quantité d'ADNc. Il existe une diversité de méthodes qui permettent une mesure complète et simultanée de l'expression de tous les gènes.

Ces méthodes peuvent être divisées en deux familles : les méthodes basées sur le principe de l'hybridation, qui permettent d'obtenir des données d'expression continues (des valeurs d'un intervalle fini ou infini) comme les méthodes q-RT-PCR et les méthodes basées sur la technologie de séquençage, qui fournissent des données d'expression discrètes (des valeurs entières) comme la technique populaire RNA-seq et la technique QuantSeq. Le tableau 3.1 résume ces quatre techniques en mentionnant quelques propriétés de chaque technique.

Techniques	Type de données	Nombre de gènes	Avec ou sans a priori
RNA-seq	données discrètes	10^4	sans
QuantSeq	données discrètes	10^4	sans
Puces ADN	données continues	10^5	avec
RT-qPCR	données continues	10^3	avec

Table 3.1: Résumé des méthodes existantes pour mesurer l'expression des gènes.

Dans cette thèse, nous nous intéressons en particulier aux données issues des techniques de séquençage RNA-seq. Une description détaillée de cette technique sera présentée dans la section suivante.

3.2 Modélisation et normalisation des données RNA-Seq

3.2.1 Le séquençage haut-débit RNA-seq

Le RNA-Seq (Cloonan and Grimmond, 2008) est une technique de séquençage à haut débit (high-throughput sequencing), appelée également séquençage de seconde génération (next-generation sequencing), qui mesure l'abondance de séquences d'ARN dans une cellule pour des milliers de gènes simultanément. Le séquençage d'un fragment d'ARN repose sur la détermination de la séquence des nucléotides d'ARN.

Les étapes du séquençage de l'ADNc par la technologie RNA-Seq sont :

1. Lecture des brins d'ADNc :
   L'ADNc est découpé en petits fragments. Ces fragments sont lus à l'aide d'un séquenceur. Les lectures qui en résultent sont appelées "reads".

2. Contrôle de qualité du séquençage ("QC") :
   Les premières paires de reads d'une lecture sont séquencées avec une grande fiabilité, mais plus on avance dans la séquence, moins elle est précise. Par la suite, les reads de mauvaises qualités sont supprimés (trimming).

3. Alignement des reads :
   L'alignement (mapping) consiste à rechercher dans le génome l'emplacement d'une sous-séquence semblable à celle de la lecture obtenue par le séquençage. Si l'on dispose du génome de référence, les lectures sont alignées avec celui-ci. Dans le meilleur des cas, le génome de référence est la séquence d'ADN complète à partir de laquelle l'ARN a été généré. En général, c'est la séquence d'ADN typique de l'espèce de l'étude. Chez *Arabidopsis thaliana* , la référence est le génome de Col-0.

4. Comptage des reads :
   Le but est de compter les reads alignés sur chaque région génomique d'intérêt. Le nombre de reads alignés sur une région d'intérêt est proportionnel au niveau d'expression de la région et à la taille de cette région. Plus le nombre moyen de reads alignés par position du génome est grand, plus le séquençage est complet. À la fin de cette étape, on connaît le nombre de reads associés à chaque gène.

La figure 3.3 décrit le schéma général de la technologie RNAseq. Comme cité dans l'article de Schutz (https://dridk.me/sc-rna-seq.html), cette figure représente le séquençage réalisé sur deux échantillons (tumeur et normal). Les ARNs sont capturés grâce à leurs queues polyA, sont convertis en ADNc puis séquencés. Les reads sont alignés sur un génome de référence afin de mesurer l'expression de chaque gène. Cette expression est proportionnelle au nombre d'ARN s'alignant sur un gène donné.

Les données obtenues à partir d'un séquençage haut-débit RNA-seq sont quantitatives. Le nombre de variables souvent le nombre de gènes, est de l'ordre de plusieurs milliers. Le plus grand désavantage de cette méthode de séquençage est que le coût d'acquisition est très élevé, même si la diminution des prix a permis de démocratiser cette méthode.

3.2.2 Modélisation statistique des données RNA-Seq

À l'issu de l'étape de séquençage, on note N_i le nombre total de reads alignés pour l'échantillon i (la profondeur du séquençage, taille de librairie, qui est une spécificité de la technologie RNA-seq). Soit Y_{ij} le nombre de reads séquencés associé à un gène j dans l'échantillon i, on a donc $N_i = \sum_j Y_{ij}$.

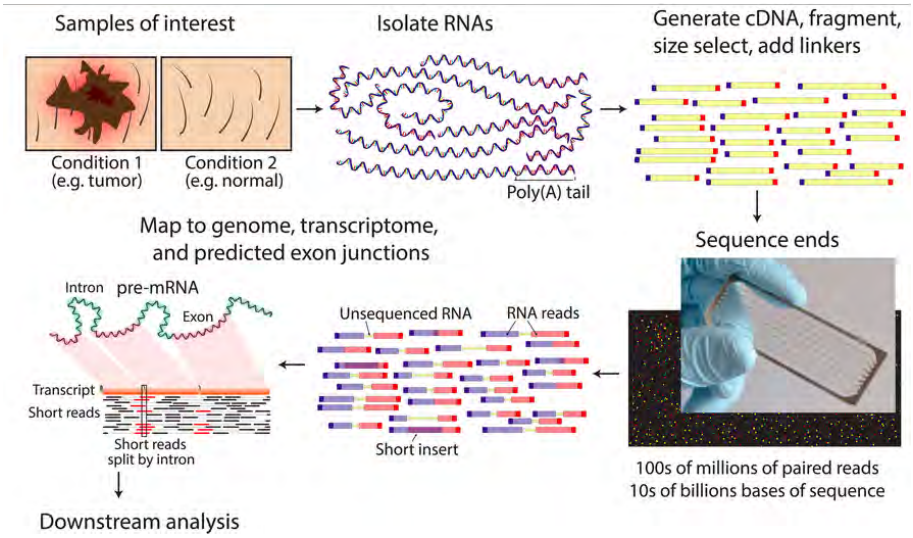


Figure 3.3: La technologie RNAseq

Modélisation des données RNA-seq

Dans les expériences de séquençage à haut débit, les données brutes consistent en des millions de reads ciblées sur des régions spécifiques du génome. Soit y_j le nombre de read séquencées pouvant être attribuées à une région particulière (c'est-à-dire un gène) j. La probabilité p_j pour un read d'être aligné sur une région j est naturellement estimée par la proportion de fragments d'ADN provenant d'une région génomique j.

Pour un gène j , son comptage Y_{ij} est modélisé par une loi binomiale de paramètre N_i et p_j :

$$p(Y_{ij} = k) = \binom{N_i}{k} p_j^k (1 - p_j)^{N_i - k}$$

Le nombre de reads N_i étant élevé et la probabilité p_j pour un read d'être aligné sur une région j étant petite, la loi binomiale de paramétres N_i et p_j peut être approximée par une loi de Poisson :

$$Y_{ij} \sim \mathcal{P}(\lambda_{ij}),$$

$$\begin{aligned} \mathrm{p}(\ \mathrm{Y}_{ij} = y_{ij}) &= \frac{\lambda_{ij}^{y_{ij}}}{y_{ij}!} e^{-\lambda_{ij}}, \\ \mathrm{E}\ (\mathrm{Y}_{ij}) &= V(Y_{ij}) = \lambda_{ij}. \end{aligned}$$

Généralement, le modèle de Poisson est utilisé lorsqu'on travaille avec des données qui contiennent des réplicats techniques d'une même condition, et donc dans le cas où on n'a pas plusieurs conditions expérimentales à étudier.

Pour les données RNA-seq, la variance d'un gène est plus élevée que sa moyenne. On parle de la surdispersion, qui est due à la variabilité biologique. Le modèle de Poisson ne peut pas être utilisé pour modéliser correctement les données d'expression RNA-Seq. C'est pourquoi, des modèles alternatifs ont été proposés pour prendre en compte cette variabilité :

Modèles de Poisson surdispersés :
Les modèles de Poisson surdispersés (Auer and Doerge,2011) se basent sur l'hypothèse d'une relation proportionnelle entre la variance et la moyenne telle que $Var(Y) = \phi E(Y)$, avec ϕ le paramètre de surdispersion.

Modèles de loi binomiales négatives :
L'avantage de l'utilisation de la loi binomiale négative (Anders and Huber,2010) est que l'on peut écrire la variance et la moyenne indépendamment. Il existe plusieurs possibilités pour définir les paramètres de la loi binomiale négative, mais l'écriture la plus courante est :

$$Y_{ij} \sim NB(\lambda_{ij}, \phi_j),$$

$$\begin{gathered} \mathrm{p}(\mathrm{Y}_{ij} = y_{ij}) = (\tfrac{\phi_j}{\phi_j+\lambda_{ij}})^{\phi_j} \tfrac{\Gamma(\phi_j+y_{ij})}{y_{ij}!\Gamma(\phi_j)} (\tfrac{\lambda_{ij}}{\phi_j+\lambda_{ij}})^{y_{ij}}, \\ \mathrm{E}(\mathrm{Y}_{ij}) = \lambda_{ij}, \\ \mathrm{V}(\mathrm{Y}_{ij}) = \lambda_{ij} + \lambda_{ij}^2 \phi_j, \end{gathered}$$

avec $\phi_j \geq 0$ le paramètre de dispersion du gène j.

Les modèles de Poisson surdispersés et les modèles de la loi binomial négatives sont considéré comme deux solutions alternatives au modèle de Poisson, permettant de prendre en compte la grande variabilité inter-échantillons observée dans les données RNA-seq. Le choix du modèle dépend de l'analyse effectuée : analyse différentielle, classification, inférence de réseaux ou intégration des données. Les paramètres doivent tenir compte du biais technique de la technologie RNA-seq. Plus d'informations sur ce sujet sont présentées dans la section suivante.

Normalisation
Le but de la normalisation est d'identifier et de supprimer les différences liées à des biais techniques entre les échantillons. L'étape de normalisation est fondamentale avant de commencer l'analyse des données de séquençage (analyse différentielle par exemple).

Prise en compte de la profondeur de séquençage par échantillon :
Pour un gène j et un échantillon donné le nombre de reads alignés est une mesure qui dépend de l'expression du gène. Ce nombre de reads est fortement dépendant de la taille de la librairie de gène N_i.

Prendre en compte le nombre total de reads alignés pour chaque échantillon (N_i) est nécessaire pour comparer les expressions de gènes de plusieurs réplicats. Le tableau 3.2 illustre ce biais. Une solution pour prendre en compte la profondeur de séquençage par échantillon consiste à utiliser le comptage par million (en anglais : count per million (Law et al.,2016)) d'un gène j pour un échantillon i notée $CPM(y_{ij})$. cette normalisation consiste à diviser le nombre de reads alignés y_{ij} par le nombre total de reads alignés N_i divisé par 10^6 :

$$CPM(y_{ij}) = \frac{y_{ij}}{N_i/10^6}$$

Comptages bruts

	gène 1	gène 2	gène 3	...	nombre total de reads
échantillon 1	80	100	170	...	2000
échantillon 2	320	150	250	...	3000

Comptages divisés par le nombre total de reads de l'échantillon

	gène 1	gène 2	gène 3	...	nombre total de reads
échantillon 1	0.04	0.05	0.085	...	2000
échantillon 2	0.106	0.05	0.083	...	3000

Table 3.2: Influence de profondeur de séquençage par échantillon sur le nombre de reads.

Prise en compte de la longueur de gène :
Pour un même niveau d'expression, un long transcrit aura plus de chances d'être séquencé et donc d'avoir plus de reads associés qu'un transcrit plus court.

Pour pouvoir comparer l'expression de différents gènes d'un échantillon donné, il est nécessaire de prendre en compte la longueur du gène sur laquelle les fragments sont alignés. La longueur d'un gène j notée L_j est exprimer en nombre de paires de bases. Pour comparer les expressions entre deux gènes, la méthode la plus simple consiste à diviser chaque comptage y_{kj} par la longueur du gène correspondant L_j. Le problème est que cette mesure ne prend pas en compte la taille de librairie N_k. Pour améliorer cette méthode de normalisation, il faut utiliser la normalisation RPKM qui prend en compte à la fois la longueur du gène et la profondeur de

séquençage (Reads Per Kilobase of exon per Million mapped Reads [Mortazavi et al., 2008]) :

$$rpkm(y_{kj}) = \frac{y_{kj}}{\frac{L_k}{10^3}\frac{N_k}{10^6}}$$

Il existe d'autres biais liée aux données RNA-seq et d'autres méthodes de normalisation. Nous présentons dans le tableau 3.3 les méthodes les plus utilisées.

Analyse différentielle

L'analyse différentielle sert à détecter les gènes différentiellement exprimés entre différentes conditions expérimentales (les gènes différentiellement exprimés entre une plante saine et une plante infectée par exemple).
Pour chaque gène j, l'analyse différentielle va déterminer si l'expression de gène j dans une première condition expérimentale est significativement différente de son expression dans une autre (ou plusieurs autres) condition expérimentale. Pour cela, des tests d'hypothèses sont utilisés.

Notons $X^k_{i,j}$ l'expression du gène j pour l'échantillon i dans la condition k. Supposons ici deux conditions expérimentales, et donc $k = 1$ ou $k = 2$. On cherche donc à savoir si, à partir des différentes expressions mesurées pour le gène j dans les conditions 1 et 2, on peut conclure que la condition expérimentale influence cette expression.

Notons m^1_j et m^2_j la moyenne d'expression du gène j dans la condition 1 et la moyenne d'expression du gène j dans la condition 2 respectivement. Le test d'hypothèse qui va nous aider à répondre à la question sera :

$$H_{0,j} = \{m^1_j = m^2_j\} \qquad contre \qquad H_{1,j} = \{m^1_j \neq m^2_j\}$$

Pour chaque gène, une statistique est calculée et est associée à une p-valeur à partir des observations. Garder ou rejeter H_0 dépend du seuil fixé α (typiquement 1% ou 5%). Si la p-valeur est inférieure à α, l'hypothèse H_0 est rejetée (le gène est différentiellement exprimé) sinon H_0 est gardée (le gène n'est pas différentiellement exprimé).
Il existe plusieurs modèles pour réaliser ces tests (un test pour chaque gène), selon le type des données de comptage et le nombre des conditions expérimentales.

Choix de modèle:

Pour établir ces tests d'hypothèses, l'expression des gènes dans chaque condition

est modélisée par une variable aléatoire et il faut effectuer un test de comparaison sur le paramètre de la distribution de la variable aléatoire pour garder ou rejeter H_0.

Il existe plusieurs modèles possibles, tels que des lois de poisson [Wang et al., 2010], des lois de poisson sur-dispersées [Auer and Doerge, 2011], des lois normales [Law et al., 2014] ou des lois binomiales négatives [Robinson and Oshlack, 2010].

L'approche la plus utilisée consiste à modéliser les données de comptage en prenant en compte leur surdispersion. La distribution binomiale négative, décrite au début de la section est la plus utilisée :

$$y_{ij}^k \sim NB(\mu_{ij}, \phi_j)$$

$$\mu_{ij} = N_i^k \lambda_i^k$$

Une fois qu'un modèle a été choisi, les paramètres du modèle correspondant doivent être estimés. Pour les modèles de distribution binomiale négative, Robinson and Oshlack [2010] et Love et al. [2014] proposent les deux méthodes les plus utilisées qui sont disponibles dans les packages R **edgeR** et **DESeq2** respectivement. Ce qui distingue ces méthodes est l'estimation du paramètre de surdispersion de la loi binomiale négative. En général, les données RNA-seq ont peu de répétitions biologiques, ce qui entraîne de mauvaises estimations des paramètres de surdispersion. Ces méthodes utilisent des techniques de partage d'informations entre les gènes pour estimer le paramètre de surdispersion de chaque gène.

Huang et al. [2015] et Costa-Silva et al. [2017] ont présenté des revues qui résument les différentes méthodes d'analyse différentielle des données d'expressions RNA-seq et les packages associés à chaque méthode.

Correction pour les tests multiples :
Pour calculer l'ensemble des gènes différentiellement exprimés, le test d'hypothèse est réalisé pour chaque gène simultanément. Il est donc nécessaire de corriger le seuil de rejet de H_0 pour contrôler correctement le nombre de faux positifs (False Discovery Rate : FDR). La solution est d'utiliser une p-value ajustée pour les test multiples. Deux grandes familles de procédures d'ajustement existent :

- La procédure de Bonferroni (Holm,1979) (FWER : Family Wise Error) : consiste à calculer la probabilité qu'il y ait au moins un faux positif dans toutes les comparaisons en calculant la probabilité d'avoir au moins un gène

non différentiellement exprimé déclaré différentiellement exprimé.

- La procédure de Benjamini-Hochberg (Hochberg and Benjamini,1990)(FDR): consiste à contrôler la quantité des faux positifs attendue dans les gènes déclarés comme différentiellement exprimés.

Le nombre de gènes différentiellement exprimés obtenus en utilisant la procédure de Bonferroni est plus faible qu'en utilisant la procédure de Benjamini-Hochberg.

Au cours de la thèse, nous, souvent, travaillons avec des données RNA-seq, pour 3 conditions expérimentales différentes. Pour déterminer les gènes différentiellement exprimés nous allons utiliser des méthodes basées sur des distributions binomiales négatives qui tiennent compte de la surdispersion. Cette analyse sera effectuée avec le package R deseq2. Une correction pour les tests multiples sera effectuée ensuite. Pour le reste de l'analyse (classification, inférence de réseau et intégration) les données seront transformées en log(CPM).

Méthode de normalisation	Description	Biais associée	Analyse adéquate
CPM (counts per million) [Law et al., 2016]	comptage mis à l'échelle par le nombre total de lectures	profondeur de séquençage	pour les comparaisons du nombre de gènes entre les répliques du même groupe d'échantillons, mais n'est pas adéquat pour les comparaisons intra-échantillon ou l'analyse DE.
TPM (transcripts per kilobase million) [Li et al., 2009]	décomptes par longueur de transcription (kb) par million de lectures cartographiés	profondeur de séquençage et longueur de gènes	pour les comparaisons de nombre de gènes au sein d'un échantillon ou entre des échantillons du même groupe d'échantillons, mais n'est pas adéquat pour l'analyse DE
RPKM [Mortazavi et al., 2008]	décomptes par longueur de transcription (kb) par million de lectures cartographiés	profondeur de séquençage et longueur de gènes	comparaisons du nombre de gènes entre les gènes d'un échantillon mais, n'est pas adéquat pour les comparaisons entre échantillons ni pour l'analyse DE
median of ratios [Love et al., 2014]	nombres divisés par des facteurs de taille spécifiques à l'échantillon déterminés par le rapport médian des nombres de gènes par rapport à la moyenne géométrique par gène	profondeur de séquençage et composition de RNA	pour la comparaison du nombre de gènes entre les échantillons et pour l'analyse DE
TMM(Trimmed mean of M values) [Robinson and Oshlack, 2010]	utilise une moyenne tronquée pondérée des rapports d'expression logarithmique entre les échantillons	profondeur de séquençage, longueur de gènes et composition de RNA	pour les comparaisons du nombre de gènes entre et au sein des échantillons et pour l'analyse DE

Table 3.3: Méthodes de normalisations connues pour les données RNA-seq et les packages R associés.

4 Inférence de réseaux de régulation de gènes

Sommaire

4.1 Les réseaux de gènes

Un graphe ou un réseau est un outil mathématique qui sert à modéliser les relations entre des variables (gènes dans notre cadre). Ces outils sont très utilisés en biologie des systèmes qui traite des interactions entre les différents acteurs au sein des systèmes biologiques. Dans sa forme, le réseau est composé d'un ensemble de nœuds représentant les entités étudiées et d'un ensemble d'arêtes modélisant un type de relations entre les noeuds (Figure 4.1). Les arêtes peuvent être orientées, elles traduisent alors l'effet d'un élément du réseau sur l'autre. Dans ce cas, on parle de graphes orientés. Dans le cas inverse, on parle des graphes non orientés. Les arêtes peuvent être parcourues d'un nœud à l'autre dans les deux sens sans savoir quel nœud affecte (agit sur) l'autre. Dans le cadre des réseaux de régulation de gènes, les nœuds représentent les gènes et les arêtes représentent les interactions entre les gènes.

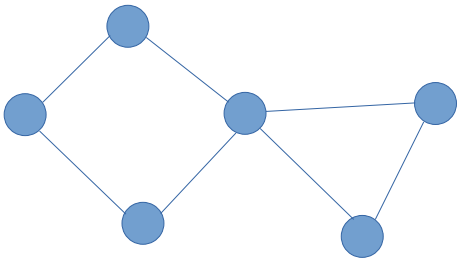


Figure 4.1: Exemple d'un réseau avec 6 nœuds (cercles) et 7 arêtes (lignes connectant deux nœuds).

Les systèmes biologiques sont souvent des processus complexes et il est nécessaire de comprendre comment les éléments qui les composent interagissent entre eux. La modélisation sous forme de réseau peut permettre d'identifier les interactions entre les composants d'un système (les gènes par exemple). Les réseaux de gènes ont prouvé leur efficacité pour saisir la complexité des interactions moléculaires sous-tendant la fonction biologique (Sanguinetti et al.,2019). L'inférence consiste à reconstruire la structure (la topologie) d'un réseau de régulation à travers l'estimation d'une matrice de coefficients (d'adjacence), qui contient des coefficients

en utilisant des modèles mathématiques. Dans ces modèles, les nœuds du réseau sont des gènes ou des groupes de gènes et les liens présentent la régulation ou la co-expression entre deux nœuds (Figure 4.2).

Au cours de l'inférence de réseau, les variables sont les différents gènes qui composent le réseau, les données sont les expressions mesurées sur ces gènes dans différentes conditions (biologiques et expérimentales). Ces mesures d'expressions peuvent être indépendantes les unes des autres ou liées chronologiquement comme le cas de mesures cinétiques d'expression. Les connexions entre les gènes permettent potentiellement ainsi d'identifier les régulateurs majeurs d'un processus biologique et de prédire les conséquences d'une perturbation d'un des éléments sur le reste du système. Cette méthode de modélisation a été utilisée pour mieux comprendre un grand nombre de maladies humaines telles que le cancer (Goh et al.,2007).

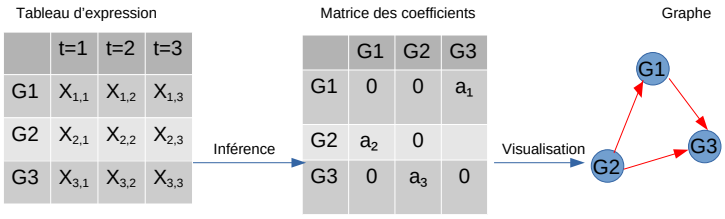


Figure 4.2: Schéma général des étapes d'inférence de réseau de gènes

4.2 méthodes d'inférence de réseaux de gènes

Il existe plusieurs méthodes pour construire des réseaux de gènes. Le choix des méthodes dépend fortement du type de données utilisées ainsi que du but de l'étude. On peut schématiquement diviser les méthodes d'inférence en deux grandes catégories : les méthodes d'inférences statiques et les méthodes d'inférences dynamiques [Sima et al., 2009, Banf and Rhee, 2017].

Les méthodes statiques tendent à déterminer toutes les interconnexions entre un ensemble de gènes en utilisant leurs expressions mesurées de manière répétée. Selon les modèles, ces connexions prennent des significations différentes, parfois

difficilement interprétables biologiquement. Parmi ces méthodes, on distingue celles dans lesquelles les interactions entre gènes sont obtenues par des tests d'indépendance ou des critères réciproques tels que les réseaux de corrélation ou de la co-expression.

On peut aussi s'intéresser à l'évolution du système à partir d'une condition donnée au cours du temps. Ceci implique le choix d'un modèle de réseau capable de caractériser la dynamique et de décrire les transitions du système dans le temps futur. Les états définis comme les valeurs du vecteur de gènes, peuvent être soit dans un domaine continu (mesure de l'expression de gènes), soit contraints à être dans un espace discret (gène différentiellement exprimé ou non). Les modèles de cette catégorie permettent la description de systèmes complexes et sont particulièrement adaptés aux données décrites sous forme de séries temporelles [Hecker et al., 2009, Penfold and Wild, 2011].

4.2.1 modèles statiques

Les méthodes d'inférence de réseaux statiques permettent de décrire les relations qui existent entre les gènes. Généralement, les données utilisées pour ces méthodes ne décrivent pas des évolutions temporelles.

Notations

Soit $G = \{1, 2, ..., p\}$ un ensemble de p gènes (p variables pour le modèle), et soit X un vecteur de $\mathbf{R}^p$, le vecteur d'expression de p gènes. La notation X_i correspond à l'expression de gène i.

Réseaux de corrélations (relevance networks)

La méthode la plus simple pour inférer un réseau de gènes est basée sur les corrélations (Butte and Kohane,1999). Cette méthode se décompose selon trois étapes :

- La première consiste à calculer la corrélation de Pearson, qui est la plus utilisée, deux à deux entre les gènes (équation 4.1).
- Fixer un seuil à partir duquel on décide si on va garder le lien ou pas dans le graphe.
- Construire le réseau : il existe un lien entre deux gènes si et seulement si la corrélation entre ces deux est supérieure au seuil fixé dans la deuxième étape.

$$cor_{i,j} = \frac{cov(X_i, X_j)}{\sigma(X_i)\sigma(X_j)} \tag{4.1}$$

L'inconvénient de cette approche est qu'elle peut mener à des fausses interprétations biologiques des relations car elle induit des liens parasites. Cette mauvaise interprétation est illustrée sur la figure 4.3

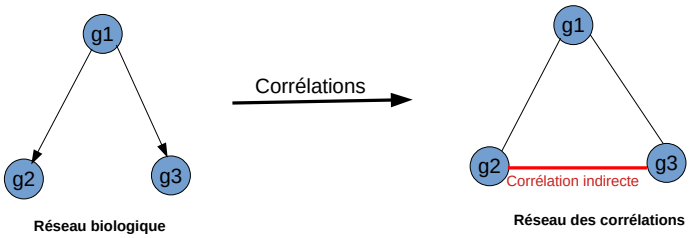


Figure 4.3: Réseau des corrélations : soient $g1$, $g2$ et $g3$ trois gènes. L'expression du gène $g1$ régule l'expression de deux gènes $g2$ et $g3$. La corrélation entre les gènes $g1$ et $g2$ est significative et la corrélation entre les gènes $g1$ et $g3$ est significative. Par la suite, il existe une forte corrélation entre $g2$ et $g3$. Le graphe produit avec des corrélations va inférer un lien entre $g3$ et $g2$, bien que ce lien direct n'existe pas dans le réseau biologique.

Une solution pour éviter ce problème est d'utiliser les corrélations partielles : les corrélations entre deux gènes sont calculées sachant l'expression de tous les autres gènes. Cette méthode va réduire le nombre des liens en gardant seulement ceux qui sont directs entre les gènes. On trouve un lien entre deux gènes (noeuds) si et seulement si les deux gènes sont dépendants conditionnellement aux autres gènes (La corrélation partielle de ces deux gènes sachant tous les autres est différente de zéro, figure 4.4). Parmi les méthodes d'inférence de réseaux basées sur les corrélations partielles, nous pouvons trouver les modèles graphiques gaussiens [Dempster, 1972].

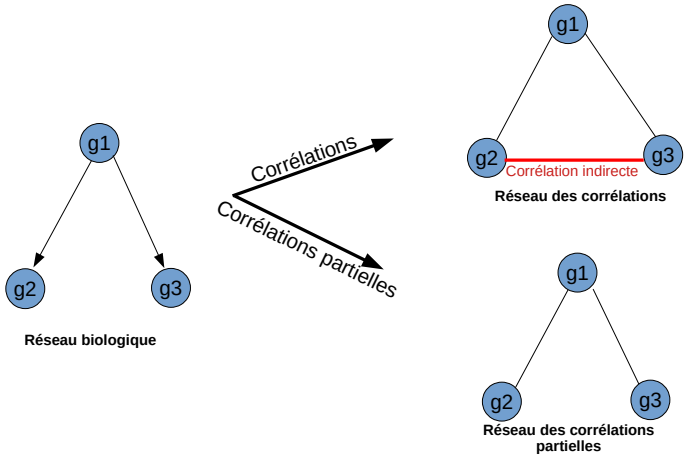


Figure 4.4: Différence entre un réseau inféré en utilisant les corrélations et un réseau inféré en utilisant les corrélations partielles.

Réseaux de corrélation partielle

Le modèle graphique Gaussien (GGM) consiste à estimer les dépendances conditionnelle entre p gènes à partir de n observations indépendantes et identiquement distribuées (idd), $(X_{i,j})$, $i = 1, \ldots, n$, $j = 1, \ldots, p$. En supposons que X suit une loi normale multivariée $X \sim N(0, \Sigma)$ où Σ est la matrice de taille $p \times p$, définie positive [Dempster, 1972].

Le modèle GGM est basé sur les résultats suivants, donnés par Dempster (1972): les variables $X_j et X'_j$ sont indépendantes conditionnellement aux autres variables si et seulement $\Sigma^{-1}_{j,j'} = 0$. La matrice de variance-covariance $\phi = \Sigma^{-1}$ est appelée la matrice de précision ou matrice de concentration. Elle est de dimension $p \times p$, ou p représente le nombre total de gènes. Cette matrice permet de décrire la dépendance conditionnelle entre les variables [Lauritzen, 1996], puisque chaque coefficient de la matrice ϕ; $\phi_{j,j'}$ $avec$ $j \neq j'$, est lié aux coefficients des corrélations partielles par la relation suivante :

$$cor(X_j \; X_{j'} | (X_k)_{k \neq j,j'}) = \frac{\phi_{j,j'}}{\sqrt{\phi_{j,j}\phi_{j',j'}}}$$

Le modèle GGM suppose que les données suivent une distribution normale multivariée. Ce modèle s'applique généralement à des données continues (tels que celles obtenus par les puces à ADN). Ce modèle ne peut pas être utilisé directement avec des données discrètes, comme les données RNA-Seq. Pour inférer des réseaux à partir de données discrètes, il est nécessaire de transformer les données afin qu'elles suivent une distribution normale multivariée. Une autre stratégie consiste à

prendre en compte la nature discrète des données RNA-Seq en utilisant un modèle basé sur des lois de Poisson [Besag, 1974, Allen and Liu, 2012].

Modèle	Avantages	Inconvénients
Corrélations	Rapide et simple. Applicable à un grand nombre de gènes	N'infère pas des relations causales. Inférence de liens non existants et non directs.
GGM	Ne nécessite aucune discrétisation des données. Ne nécessite pas d'informations préalables.	Inadapté aux gros réseaux ou réseaux très connectés. N'infère pas des relations causales. N'infère pas d'interaction indirecte à partir des variables d'état cachées.

Table 4.1: Résumé des propriétés, avantages et inconvénients des réseaux obtenus en utilisant les corrélations et les corrélations partielles.

4.2.2 modèles dynamiques

Les processus biologiques sont dynamiques et peuvent être décrits par des séries temporelles. Les réseaux de gènes jouent un rôle important dans la modélisation mathématique de ces processus. Dans cette section, nous introduisons trois différentes méthodes largement utilisées pour inférer la structure d'un réseau à partir de données dynamiques.

Notations

Soit $G = \{1, 2, ..., p\}$ un ensemble de p gènes (p variables pour le modèle), et soit X^t un vecteur de $\mathbf{R}^p$, le vecteur d'expression de p gènes au temps t, avec $t \in [0, T]$. On peut écrire X^t comme : $X^t = \{x_1^t, ..., x_p^t\}$.

Modèles à temps discret

Sous l'hypothèse Markovienne classique, l'expression d'un gène i au temps t dans le réseau s'exprime en fonction de l'expression de ces p gènes au temps précédant, et une forme simple de modélisation est alors

$$X^t = F(X^{t-1}) + \epsilon_t, \tag{4.2}$$

où F représente une fonction de $\mathbf{R}^p$ dans $\mathbf{R}^p$ et ϵ_t va décrire le bruit gaussien associé aux mesures de l'expression au temps t.

Lorsque les états X^t sont discrets, on peut considérer une version probabiliste de ce modèle sous forme d'une chaîne de Markov $p(X^t \mid X^{t-1})$ [Dewey and Galas, 2001], [Dewey, 2002], qui vérifie alors pour toute trajectoire $(X^0, ..., X^T)$

$$p(X^0, ..., X^T) = p(X^0) \prod_{t=1}^{T} p(X^t \mid X^{t-1}). \tag{4.3}$$

Le modèle 4.2 peut se décomposer en p modèles unidimensionnels pour chaque gène :

$$x_i^t = f_i(x_1^{t-1}, \ldots, x_p^{t-1}) + \epsilon_i^t \tag{4.4}$$

avec x_i^t l'expression d'un gène i au temps t, f_i une fonction qui permet d'obtenir l'expression de gène i et ϵ_i^t est un bruit de mesure de l'expression i au temps t.

Les fonctions f_i peuvent avoir plusieurs formes, la plus simple étant la forme linéaire. Le modèle 4.2 peut alors s'écrire :

$$X^t = A\, X^{t-1} + \epsilon^t, \tag{4.5}$$

avec X^t le vecteur de l'expression de tous les gène au temps t, A la matrice de transition et ϵ^t le vecteur de bruit gaussien. La détermination de la matrice d'adjacence A donne la structure de graphe du réseau de gènes.

Modèles à temps Continu

Les équations différentielles ordinaires (Ordinary Differential Equation : ODE) fournissent une description de la dynamique du système en reliant la dérivée temporelle d'une variable à sa valeur [Bansal et al., 2006] :

$$\frac{dX}{dt} = f(X, \theta, u(t), t), \tag{4.6}$$

où $X(t) = X^t$ est le vecteur d'expression de l'ensemble des gènes à modéliser. Les interactions entre les gènes sont décrites dans les paramètres θ, f représente la fonction qui décrit l'évolution de l'expression des gènes et $u(t)$ représente une perturbation potentielle appliquée, qui peut être exprimée explicitement ou implicitement selon le choix de la fonction f.

La construction de réseau de régulation en utilisant l'équation différentielle consiste à estimer les paramètre θ d'interaction, par exemple une matrice d'interaction.

On peut distinguer deux principales familles de formes pour f : linéaire ou non linéaire.

Il existe des systèmes biologiques non linéaires tels que les rythmes circadiens

[Goldbeter, 1995] et donc dans ce cas, on va être placé dans le cadre des équations différentielles non linéaires (Heinrich and Schuster,2012). L'estimation des paramètres de ces équations va être plus difficile par rapport aux équations linéaires, et c'est pour cela que généralement ces modèles sont utilisés avec un petit nombre des gènes.

Dans le cas des équations différentielles linéaires, on peut classiquement écrire le modèle 4.6 sous la forme :

$$\frac{dX}{dt} = A\,X, \tag{4.7}$$

où X représente le vecteur d'expression des gènes et A représente la matrice des coefficients qui décrit la relation entre les différents gènes et de dimension $p \times p$.

L'équation 4.7 s'écrit en explicitant le produit matriciel :

$$\frac{dX_i}{dt} = \sum_{j=1}^{p} a_{i,j}\ X_j \tag{4.8}$$

Avec i et j deux gènes de G, $X_i(t)$ le vecteur d'expression du gène i au temps t et $a_{i,j}$ l'impact du gène j sur le gène i.

Les deux modèles 4.7 et 4.8 affirment que l'expression du gène i au cours du temps dépend de la somme de l'expression de tous les autres gènes pondérées par les coefficients $a_{i,j}$ qui décrivent l'impact des gènes j sur le gène i.

On peut enrichir ce modèle (4.8) en ajoutant des paramètres relatifs à l'impact d'une perturbation externe sur le système de l'organisme étudié. On obtient le modèle suivant :

$$\frac{dX_i}{dt} = \sum_{j=1}^{p} a_{i,j}\ X_j + b_i.u, \tag{4.9}$$

avec u la perturbation appliquée au système biologique et b_i l'action de la perturbation sur le gène i.

Il est classique alors de simplifier ce modèle 4.9 en discrétisant le temps, en approchant le terme de dérivation par la différence finie :

$$\frac{X_i^{\delta t+t} - X_i^t}{\delta t} = \sum_{j=1}^{p} a_{i,j}\ X_j^t + b_i.u^t \tag{4.10}$$

On obtient alors une forme équivalente au modèle markovien à temps discret, avec une réécriture de la matrice A en $I + \delta t A$. Ces modèles peuvent également être vus comme des modèles autorégressifs d'ordre 1 (VAR(1)).

4.2.3 Méthodes d'inférence de graphe

Tous les modèles présentés peuvent être transformés en des modèles de régression linéaires classiques pour chaque gène, en régressant chaque gènes au temps t sur les expressions au temps $t-1$ de l'ensemble des autres gènes :

$$Z = Y\beta + E, \tag{4.11}$$

Où $Z = [X_2, ..., X_T]'$ est le vecteur réponse de taille $T-1 \times p$, β de dimension $p \times p$ la transposée de la matrice des coefficients, $Y = [X_1, ..., X_{t-1}]'$ et $E = [\epsilon_2, ..., \epsilon_T]'$ de taille $T-1 \times p$.

Pour déterminer β, ce modèle peut être estimé par la méthode de Moindres Carrés Ordinaires (OLS : ordinary least squares). Dans ce cas $\hat{\beta}$ va être déterminée comme :

$$\hat{\beta} = \underset{\beta}{argmin}(\frac{1}{n}||Z - X\beta|| \tag{4.12}$$

Cette méthode d'estimation conduit généralement à une matrice des coefficients ne contenant que peu de coefficients nuls.

Les méthodes de pénalisation comme la méthode LASSO (Least Absolute Shrinkage and Selection Operator) permettent au contraire de rechercher des matrices de coefficients qui soient creuses (avec de nombreux coefficients nuls). Dans ce cas $\hat{\beta}$ va dans ce cas être déterminée comme :

$$\hat{\beta} = \underset{\beta}{argmin}(\frac{1}{n}||Z - X\beta|| + \lambda||\beta||_1) \tag{4.13}$$

Avec λ est le paramètre de pénalisation.

Dans le cas des modèles autorégressif d'ordre 1 (VAR(1)), modèles markoviens à temps discrets où ϵ_t suit une loi normale de moyenne nulle et de matrice de covariance fixée, le package R SIMONE (Statistical Inference for MOdular Network) propose en particulier une méthode d'inférence basée sur une telle pénalisation qui permet de trouver des hubs lorsqu'on infère des réseaux de régulations dynamiques de gènes. Une description complète de la méthode d'inférence est donnée dans l'article de Charbonnier et al. [2010].

Modèle	Avantages	Inconvénients
Chaînes de Markov	Des modèles probabilistes	Inadapté à un grand nombre de gènes. Coût computationnel fort.
ODE	Convient de manière fiable aux données temporelles pour un petit nombre de gènes et de conditions. Simplicité d'implémentation	Inadapté à un grand nombre de gènes et inadapté aux modèles avec de nombreux paramètres [Sachs et al., 2005]
Var(1)	décrit explicitement les relations causales. Simplicité d'implémentation	Les points de mesures doivent être régulièrement espacées. Le nombre des variables doit être au plus du même ordre que le nombre de mesures temporelles.

Table 4.2: Résumé des propriétés, avantages et inconvénients des réseaux obtenus en utilisant les modèles de Markov, les équations différentielles et les modèles autorégressif d'ordre 1.

4.3 Classification

L'inférence de réseau de gènes est un problème complexe à cause du faible nombre d'observations (expressions de gènes) par rapport au nombre de variables (gènes). Pour résoudre ce problème, il est nécessaire de réduire le nombre de variables avant l'inférence. Pour réduire la dimension (le nombre de gènes), on peut utiliser des connaissances biologiques a priori (issues d'expérience biologique ou d'utilisation des informations déjà existante) en choisissant un sous ensemble de gènes d'intérêt par exemple. Une deuxième méthode, sans a priori, consiste à utiliser des méthodes de classification, qui servent à grouper les gènes en se basant sur leurs expressions.

Dans cette thèse, nous avons opté pour des approches sans a priori en utilisant des méthodes de classification.

Pour classifier les données RNA-seq, on peut soit transformer les données pour utiliser des méthodes de classification classiques ([Severin et al., 2010], [Anders and Huber, 2010]), soit utiliser un modèle adapté à ce type des données RNA-seq [Witten, 2011]. Dans l'article de Anders and Huber [2010], les données RNA-seq ont été transformées en utilisant la méthode VST (Variance stabilizing transformation) pour stabiliser la variance puis une classification hiérarchique (en utilisant la distance euclidienne) a été appliquée sur les données transformées. Une transformation en RPKM (Reads Per Kilobase of exon per Million mapped Reads) puis une classification hiérarchique en utilisant la corrélation de Pearson a été appliquée dans l'étude de Severin et al. [2010]. Pour l'étude de Witten [2011], une distance entre les réplicats biologiques calculée en utilisant un modèle de Poisson log linéaire a été utilisée et puis une classification hiérarchique a été appliquée.

Les méthodes de classification des séries temporelles sont des techniques d'exploration de données visant à former des groupes homogènes de formes sans connaissance préalable des groupes. Les classes sont formées en regroupant les séries temporelles qui ont une similarité maximale avec les éléments du même groupe et une similarité plus faible avec des objets d'autres groupes.

Plusieurs méthodes de regroupement de séries temporelles existent :
Les méthodes basé sur la similarité (similarity-based) : [Sakoe and Chiba, 1978], [Lines and Bagnall, 2015]
La majorité des travaux sur la classification des séries temporelles sont basés sur des mesures de distance qui vont être utilisées pour grouper les séries temporelles. Ces mesures de distance sont presque exclusivement évaluées à l'aide d'un classificateur du plus proche voisin. Les mesures de distance de référence sont la distance euclidienne (ED) et la déformation temporelle dynamique (dynamic time warping : DTW).

Les méthodes basées sur un intervalle (interval-based): [Rodríguez and Alonso, 2004]
C'est sont des méthodes qui décrivent les les caractéristiques des intervalles de chaque série temporelle. Pour les séries de longueur T, il y a $T(T-2)/2$ intervalles contigus possibles. Les deux questions principales posées avant d'utiliser ces méthodes sont : comment gérer augmentation importante de la dimension de l'espace des caractéristiques, et que faire réellement de chaque intervalle (Bagnall et al.,2017) ? Une méthode propose par exemple de n'utiliser que des intervalles de longueurs égales à des puissances de deux (Rodríguez and Alonso,2004).

Ces méthodes ont montré leur efficacité pour la classification de séries temporelles

longues et régulièrement espacées. Comme nous le verrons dans le prochain chapitre, ces méthodes sont peu intéressantes pour les séries temporelles courtes comme celle associées aux données RNA-seq.

4.4 Évaluation de la qualité de la classification

Au delà des critères descriptifs (le nombre des classes, taille de chaque classe, etc.), il existe des critères qualitatifs permettant de décrire la classification réalisée. Ces critères sont :

- La distance intra-classes (homogénéité d'une classe) : cette distance se base sur le calcul de la somme des carrés des distances de chaque élément au centre de la classe à laquelle il appartient.
- La distance inter-classes (séparation entre classe) : La la distance entre les centres des différentes classes.

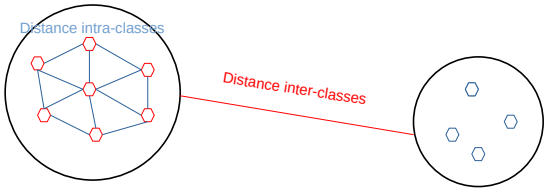


Figure 4.5: Schématisation des distances inter et intra classes

- L'indice de silhouette : cet indice permet d'évaluer si ce point appartient au "bon " groupe (classe). L'indice est donné par la différence entre la distance moyenne entre les éléments du même groupe et la distance moyenne avec les éléments des autres groupes voisins. L'indice de silhouette varie entre -1 (pire classification) et 1 (meilleure classification).

- L'indice de Davies-Bouldin : L'idée de cet indice est de comparer les distances intra-classe (l'homogénéité), que l'on veut faibles, aux distances inter- (la séparation), que l'on veut grandes. Cet indice est donné par la moyenne du rapport maximal entre la distance d'un point au centre de son groupe et la distance entre deux centres de groupes, et varie entre 0 (meilleure classification) et $+\infty$ (pire classification).

4.5 Des exemples d'inférence de réseau pour les données d'expression de gènes

Plusieurs travaux de biologie végétale et humaine utilisant des modèles dynamiques d'inférence de réseau ont été publiés depuis 10 ans.

Dans l'article de Guthke et al. [2005], Un modèle d'équation différentielle linéaire est utilisé pour comprendre la réponse immunitaire humaine lors d'une infection bactérienne. Pour pouvoir appliquer ce modèle, ils réalisent une classification. Le nombre initial de gènes était 7619, pour 5 points de mesures temporelles. Ce nombre a été ramené à 6 par une approche de classification, pour obtenir un réseau qui contient donc 6 noeuds.

Un réseau booléen dynamique à été utilisé dans l'article Martin et al. [2007], après une classification par kmeans. Les 34000 gènes (et 12 points temporels) ont été groupés en 12 classes pour obtenir un réseau de 12 noeuds. Un ensemble d'équations différentielles non linéaires est utilisé par Kimura et al. [2005] pour inférer un réseau de régulation. Le modèle a été appliqué après une classification hiérarchique classique sur l'expression temporelle de 612 gènes (14 points temporelles). Ces expressions ont été groupées en 25 classes pour obtenir un réseau de 25 noeuds. Pour prédire l'expression globale des gènes de *Halobacterium* sous de nouvelles perturbations, Bonneau et al. [2006] ont classifié l'expression statique d'environ 2400 gènes en 531 classes et ont utilisé des équations différentielles pour inférer le réseau.

En 2011, Yao et al (2011) ont inféré un réseau dynamique pour le contrôle de l'acclimatation à un changement de lumière chez *A. thaliana*. Les expressions des gènes ont été mesurées avec des puces ADN à 0h, 0.5h, 2h, 8h et 48h après un changement de qualité de lumière. 48h après le changement de lumière, plus de 2000 gènes ont été déclarés différentiellement exprimés. Un important réseau de régulation est donc mis en place lors de l'acclimatation des plantes à un changement de lumière. L'utilisation d'un modèle d'inférence de réseau était impossible vu

le grand nombre de gènes différentiellement exprimés. La modélisation a donc nécessité plusieurs étapes préalables afin de réduire ce nombre. La première étape a consisté à créer une liste de gènes impliqués dans l'acclimatation à la lumière via une fouille de données de la littérature et de bases de données. Pour chacun des gènes de cette liste, une étude des cis-motif présents sur leur promoteur a été effectuée. Cette étude a permis de créer pour chaque gène une liste de facteurs de transcription et une liste de régulateurs potentiels de ces gènes. La deuxième étape a consisté à ne conserver parmi les régulateurs potentiels que ceux différentiellement exprimés dans les expériences de puce ADN, et à étudier la corrélation entre l'expression de chaque gène (données de puce) et l'expression de ces régulateurs. Si la corrélation entre le gène et les facteurs de transcription était faible alors ils étaient retirés des deux listes. À la fin de ces deux étapes, un premier réseau de 65 gènes dont 7 facteurs de transcription ont été inférés. La dernière étape, a consisté à inférer un réseau en utilisant les équations différentielles à partir du premier réseau inféré et des données de puce ADN. Cette modélisation dynamique de régulation via les équations différentielles a permis de mettre en évidence le rôle central de deux facteurs de transcription (HB-1 et RR10) dans la réponse permettant l'acclimatation à un changement de lumière.

Pour comprendre le système de défense d'*A. thaliana* contre le champignon *B.cineria*, Windram et al (2012), ont inféré un réseau dynamique des gènes. Les expressions des gènes ont été mesurées avec des puces ADN toutes les 2h entre 2h et 48h après l'infection à *B.cineria*. 9838 gènes ont été déclarés différentiellement exprimés. Pour réduire les dimensions du problème, les auteurs de ce papier ont groupé les gènes ayant des expressions temporelles proches dans une même classe. L'algorithme de classification utilisé dans ce papier est le SplineCluster (Heard et al.,2005). En utilisant cet algorithme, 44 classes ont été constituées, ce qui a permis la réduction significative de la dimension du problème. Pour l'inférence du réseau, les moyennes de chaque classe ont été considérées. Le réseau inféré contenait 44 noeuds qui représentaient les classes des gènes ainsi qu'un noeud supplémentaire représentant la croissance du phénotype (Figure 4.6.) L'analyse du réseau obtenu a permis de montrer le rôle important du gène TGA3 dans la défense contre les agents pathogènes nécrotrophes.

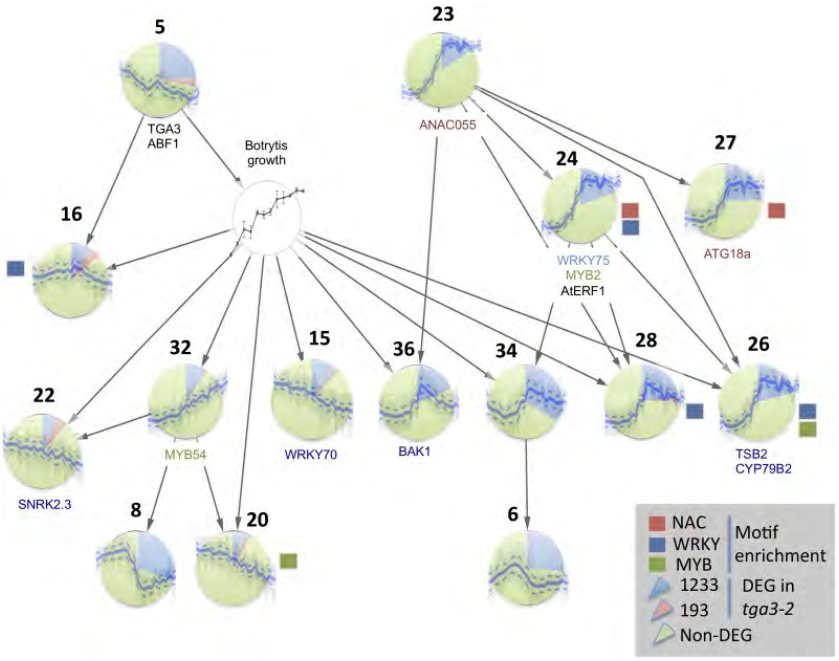


Figure 4.6: Réseau inféré pour décrire la réponse immunitaire de *A. thaliana* lors d'une infection à *B. cinerea*. Les nœuds numérotés représentent les clusters de gènes exprimés de manière différentielle lors d'une infection à *B.cineria*. La couleur des nœuds indique la proportion de gènes de cluster.

En 2022, Deshpande et al. ont développé SINGE (single-cell inference of networks using Granger ensemble), un algorithme utilisant la régression de causalité de Granger basée sur le noyau pour lisser les pseudo-temps irréguliers et les valeurs d'expressions manquantes. Cet algorithme agrège les prédictions d'un ensemble d'analyses de régression pour compiler une liste classée des interactions candidates entre les régulateurs transcriptionnels et les gènes cibles. L'algorithme a été appliqué à un ensemble de données qui décrit la différenciation cellulaire induite par l'acide rétinoïque des ESC de souris sur 96h. Le but de cet algorithme était d'améliorer la prédiction en utilisant des modèles causaux sur des données temporelles.

Les réseaux de régulations basés sur la corrélation ne décrivent que des coexpressions simultanées de gènes, mais ne décrivent pas la dynamique des données. Pour résoudre ces limitations et prendre en compte la variation temporelle des données Specht and Li (2017) ont développé une méthode LEAP (Lag-based Expression Association for Pseudotime-series) capable de capturer les associations masquées par les décalages temporels en utilisant les informations temporelles disponibles dans les données scRNA-Seq (Single-cell RNA sequencing). Cette méthode est publiée sous forme d'un package R LEAP. LEAP exige que les cellules soient ordonnées le

long d'une trajectoire pseudo-temporelle et calcule la corrélation par paires entre les gènes i et$j$ à différents retards le long de cette trajectoire. L'algorithme choisit une taille de fenêtre s, et pour tous les instants de départ t et retards l possibles, calcule le coefficient de corrélation de Pearson entre les s premières observations de i commençant à l'instant t et les observations de j commençant à l'instant $t + l$. Le score pour une relation de régulation de i à j est le maximum de tous les coefficients de corrélation calculés.

SCODE est un algorithme basé sur les équations différentielles ordinaire et nécessite un pseudo-temps comme entrée en plus de la matrice d'expression pour apprendre le réseau de régulation [Matsumoto et al., 2017]. SCODE utilise une représentation dimensionnelle inférieure de la dynamique d'expression majeure des régulateurs, et utilise une matrice de transformation linéaire, pour estimer la dynamique d'expression observée de tous les gènes à partir de l'expression majeure de régulateurs. Le réseau de régulation des gènes est représenté sous forme d'une matrice et estimé en résolvant un ensemble de régressions linéaires.

Enfin, un article récent de McCalla et al. (2023), présente l'ensemble des algorithmes assez récents d'inférence de réseau sur les données RNA-seq [McCalla et al., 2023]. La figure 4.7 de l'article présente les données et les algorithmes utilisés ainsi que les méthodes d'évaluation de ces différentes méthodes.

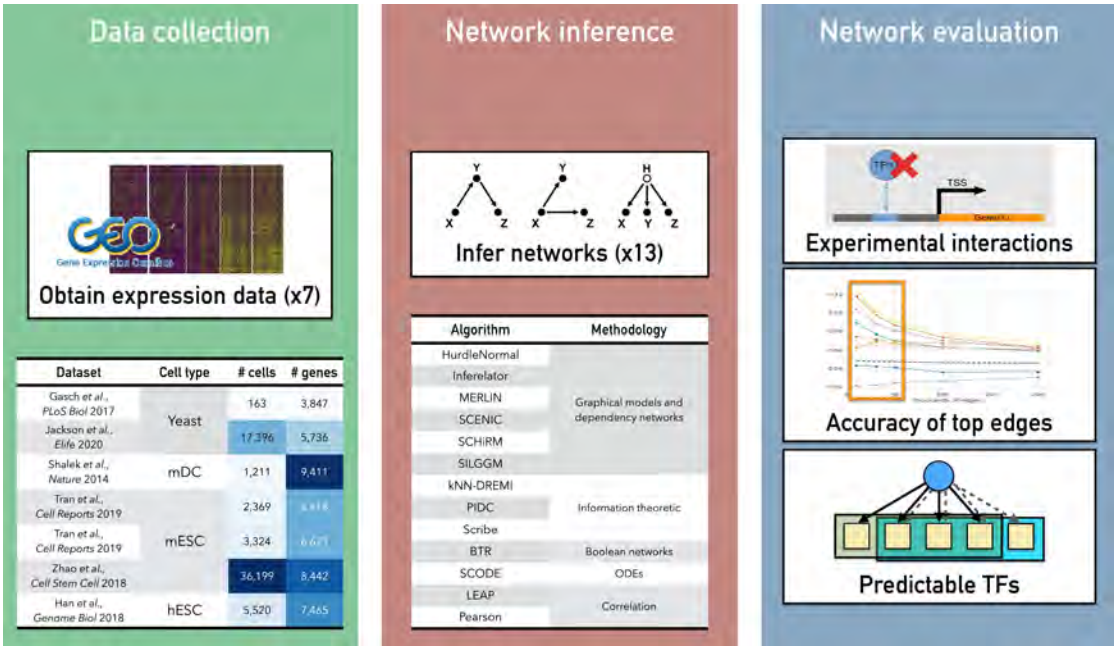


Figure 4.7: Aperçu de l'étude comparative d'inférence de réseau présentée dans l'article de McCalla et al. (2023) : 13 algorithmes développés ont été étudiés sur des jeux de données issus de trois espèces différentes : levure, souris et humain.

5 Acquisition des données RNA-seq et hypothèses sous-jacentes

Sommaire

Afin d'identifier les effets moléculaires causés par les stimulations sonores, nous avons mis en place une approche basée sur l'exploitation de données de RNA-seq. Cette approche a un double objectif :

- identifier les gènes significativement modulés par les stimulations acoustiques chez les plantes saines afin d'identifier.
  - les gènes de réponse aux stimulations, c'est à dire les gènes modulés avec la même amplitude quel que soit le nombre des stimulations acoustiques.

  - les gènes impliqués dans la mémoire transcriptionnelle, c'est à dire les gènes modulés differentiellement après chaque stimulation [Avramova, 2019].

- identifier chez les plantes exposées au son avant l'infection les gènes participant au priming, c'est à dire les gènes exprimés plus fortement en valeur absolue.

Il est a noter que pendant cette thèse nous nous sommes focalisés sur les effets des stimulations sonores sur la plante. Les données recueillies permettraient cependant de décrire également comment la virulence du champignon est modulée par la plante primée ou comment les stimulations modulent l'interaction plante-champignon en considérant les transcrits de l'holobionte.

5.1 Réflexions concernant le protocole expérimental

5.1.1 Hypothèse centrale du travail

L'objectif est de construire l'évolution de l'expression des gènes en fonction du nombre de stimulations acoustiques. Pour des raisons pratiques, les stimulations sonores n'ont lieu qu'une fois par jour ce qui établit une correspondance simple entre le nombre de stimulations et le temps. Cela contraint également à 24h la résolution temporelle à laquelle nous étudions l'effet des ondes acoustiques sur l'expression des gènes. Bien que le RNA-seq ce soit démocratisé, ce type de mesure reste relativement coûteux. Afin de limiter le nombre d'échantillons nous faisons l'hypothèse que la modulation de l'expression des gènes par les stimulations acoustiques est plus importante que les fluctuations d'expressions des gènes chez les plantes non stimulées [I]. Cette hypothèse est vérifiée si, en l'absence de sollicitations extérieures, les expressions des gènes sont constantes ou suivent une dynamique temporelle de période maximale 24h ou d'un de ses sous multiples.

Cette hypothèse est justifiée d'un point de vue biologique par le fait :

- que l'horloge circadienne est en grande partie régulée par la photopériode [Takahashi et al., 2015] et que la photo-période est contrôlée en laboratoire.

- qu'à 4 semaines les mécanismes de croissance sont limités chez *A. thaliana* laissant supposer que peu de fluctuations d'expressions soient visibles en RNA-seq.

[I] Cette hypothèse a été vérifiée depuis dans l'équipe QIP (Florend Delplace, Communication personnelle).

Cette hypothèse permet de proposer un protocole dans lequel le prélèvement des RNAs des plantes stimulées et contrôles se fait le même jour. Le nombre d'échantillons contrôle s'en trouve ainsi réduit (fig. 5.1).

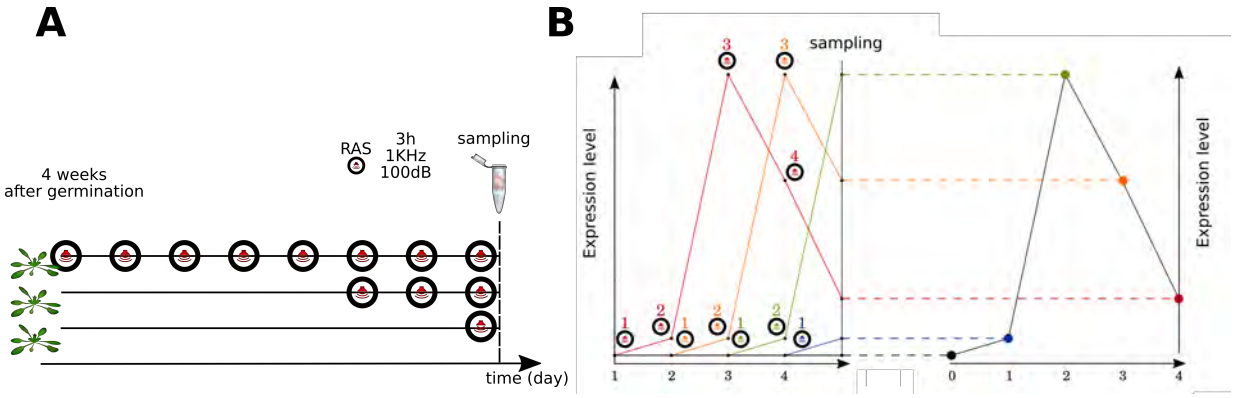


Figure 5.1: Protocole d'échantillonnage expérimental et reconstruction de la dynamique de l'expression génique. A. Protocole d'échantillonnage : les plantes ont été exposées à 0 à 8 stimuli acoustiques (1 KHz, 100 dB, 3h/jour). Des échantillons de feuilles ont été extraits en même temps pour toutes les plantes. B. Le protocole d'échantillonnage expérimental permet la reconstruction de la dynamique de l'expression génique en supposant que l'expression génique ne dépend que du nombre de stimuli acoustiques.

5.1.2 Choix de la fréquence et de l'intensité des ondes acoustiques

Une onde acoustique est caractérisée par sa fréquence et son intensité. Le choix de ses paramètres doit obéir à certaines contraintes techniques, biologiques et expérimentales. De manière triviale, l'intensité du son ne peut être inférieure à l'intensité du son en chambre de culture qui est de l'ordre de 70 dB. Pour les fréquences, on peut distinguer deux types de fréquences :

- les fréquences de la base modale qui créent des déplacements importants qui ne dépendent que des propriétés géométriques et mécaniques des feuilles.
- les fréquences quelconques.

Au delà de considérations techniques, le choix d'une fréquence propre est peu adapté à l'étude de l'effet du son sur la réponse de la plante au son dans la mesure où les déplacements induits ne dépendent plus de la fréquence de sollicitations. Le choix d'une fréquence propre est probablement un choix plus pertinent pour l'étude de la proprioception [Moulia et al., 2021]. Les modes propres les plus importants énérgétiquement sont observés à basse fréquence ($\sim$ 10 Hz). A partir

de 100 Hz, on observe des modes d'ordre supérieur induisant des champs de déplacements complexes combinaisons linéaires très amorties de modes plus basses fréquences. L'ordre de grandeur de l'amplitude des déplacements induits est de 100 nm (épaisseur d'un cheveu) pour des fréquences inférieures au kHz et 10 nm (épaisseur d'un globule rouge) au delà.

Les travaux publiés [Ghosh et al., 2016] se sont intéressés à des fréquences supérieures à 200 Hz. Ils montrent que les gènes modulés dépendent de la fréquence de l'onde acoustique. Cependant, aucune relation avec le phénotype de résistance n'est connue. Quelques expériences de vérification menées au laboratoire testant les fréquences de 1 kHz et 2 kHz ont permis de montrer que, bien qu'associée à des paysages transcriptomiques légèrement différents, les deux fréquences ont un effet significatif sur le phénotype de résistance.

Les travaux les plus détaillés sur les impacts physiologiques des ondes acoustiques sur la plante ont été menés à la fréquence de 1 kHz. Par exemple, (2017a), ont rapporté l'effet de la répétition de 10 stimulations sonores de 3h par jour à 1 kHz, 100 dB sur la production d'acide salicylique et jasmonique. Le niveau de SA dans les plantes primées par le son est alors augmenté par rapport aux plantes contrôle.

Notre choix de fréquence s'est ainsi porté sur la fréquence la plus utilisée dans les publications : 1 kHz. Cette fréquence est également la fréquence utilisée pour l'étude biophysique des vibrations de tympans animaux. Une partie du travail concernant le son et les plantes au laboratoire consiste à comprendre la perception des vibrations induites par le son dans les feuilles. Ce travail se fait à la fréquence de 1 kHz de sorte à pouvoir comparer les résultats obtenus sur feuilles à ceux obtenus sur les tympans (autre membrane biologique). Établir des résultats à 1 kHz offre donc aussi la possibilité d'une compréhension plus ample de la perception des ondes acoustiques par les plantes.

En ce qui concerne le choix de l'intensité ; on trouve deux intensités dans la littérature : 100 dB et 80 dB. Nous avons choisi de travailler à 100 dB pour être certain que l'intensité de l'onde acoustique soit plusieurs ordres de grandeur supérieure au bruit ambiant des chambres de cultures. On peut noter que 100 dB correspondent à une pression dynamique de 1 Pa soit 10^{-5} fois la pression atmosphérique.

Cette analyse nous conduit au protocole expérimental suivant.

5.2 Protocole expérimental

5.2.1 Matériel végétal pour le RNA-seq

Les plantes d'*A. thaliana* Col-0 ont été cultivées en chambre de culture à température et humidité constantes (23°C, 80% HR) avec 9 h/jour de lumière (180 µmol/m2). Après 3 semaines, les plantes ont été exposées à des stimulations acoustiques répétées (RAS pour Repeated Acoustic Stimuli) de fréquence 1 kHz, et d'intensité 100 dB. Les stimulations sont appliquées 3 h/jour pendant 0 à 8 jours au début de la période de jour. Une première partie des échantillons est prélevée à la fin de la dernière stimulation afin de déterminer l'effet du son sur le transcriptome. Chaque échantillon correspond à des tissus foliaires d'une même plante. Les feuilles des plantes restantes ont été infectées avec le champignon nécrotrophe *S. sclerotiorum* après la dernière stimulation acoustique. L'infection consiste au dépôt d'un plug de PDA (Dextrose de pomme de terre en milieu agar) contenant de mycléium sur une feuille de la plante en pot. Cette méthode permet d'identifier clairement le départ de l'infection. 24h après l'inoculation, on prélève les échantillons dans la zone périphérique de l'infection entourant la nécrose [Peyraud et al., 2019]. Le nombre d'échantillons prélevée par le RNAseq par modalité est récapitulé dans la table Table 5.1.

Nombre de stimuli 3h 1KHz, 1Pa	0		1		3		8	
Infection	oui	non	oui	non	oui	non	oui	non
Replicats biologiques	4	4	4	4	3	3	4	4

Table 5.1: Nombre d'échantillon analysé en RNAseq par modalité.

5.2.2 Évaluation de l'impact des stimulations acoustiques répétées sur le phénotype de résistance

Pour déterminer phénotypiquement la résistance des plantes primées avec les mêmes modalités, nous avons utilisé d'autres plantes âgées de 4 semaines. Le phénotypage se fait sur feuilles détachées en suivant une méthode développée et utilisée par l'équipe QiP [Barbacci et al., 2020]. Cette méthode permet de prendre toutes les 10 minutes une photographie haute résolution de l'ensemble des feuilles pendant toute la durée de l'infection qui a lieu dans une enceinte à forte humidité et à température fixe de 23°C. Un programme d'analyse d'image permet d'extraire, feuille à feuille et image par image la nécrose de la partie verte de la feuille (https://github.com/A02l01/INFEST//master/readme.md). La progression

cinétique de la feuille est une courbe ressemblant à une sigmoide c'est à dire composée

- d'une phase de latence pendant laquelle la maladie n'est pas visible. Cette phase correspond au temps qu'il faut au champignon pour pénétrer dans les tissus végétaux et initier sa croissance.
- une phase de croissance exponentielle des symptômes.
- une phase asymptotique correspondant au moment où le champignon a colonisé l'ensemble de la feuille.

Les travaux antérieurs de l'équipe QiP ont montré que la phase asymptomatique était suffisante pour décrire la résistance de la plante [Barbacci et al., 2020]. Contrairement à la publication qui prenait comme indicateur de résistance le temps caractéristique de doublement de lésion, nous avons utilisé le taux de croissance de la partie exponentielle. Ce taux de croissance est estimé dans l'espace log par la librairie R segmented [Muggeo et al., 2008]. Plus le taux de croissance est élevé plus la susceptibilité (l'inverse de la résistance) à la maladie est forte.

5.3 Cohérence des données

Une première vérification classique des résultats du RNA-seq se fait par une ACP pour attester de la cohérence des données.

La figure 5.2.A présente la projection des répétitions biologiques des ARN associées aux plantes saines exposées sur les deux premiers axes de l'ACP. Comme attendu, les points sont groupés par durée d'exposition au son. 36% de la variance est expliquée par le premier axe et 20% par le deuxième axe. La même démarche a été appliquée pour les données des plantes primées puis infectées (Figure.5.2.B). 25% de la variance était expliquée par le premier axe et 22% sur par le deuxième axe. La figure 5.2.C présente l'ensemble de toutes les modalités (plante saine exposés au son et plante exposée au son puis infecté). Les différentes combinaisons des modalités testées sont parfaitement séparées. L'axe 1 explique la variance liée à l'infection alors que l'axe 2 explique les variations induites par les stimuli acoustiques chez les plantes saines.

La figure 5.3 présente les heatmap des top 1000 gènes exprimés dans les différents traitements. La classification hiérarchique montre également que l'infection est le facteur le plus discriminant.
L'ACP et le heatmap montrent la cohérence du jeu de données constitué.

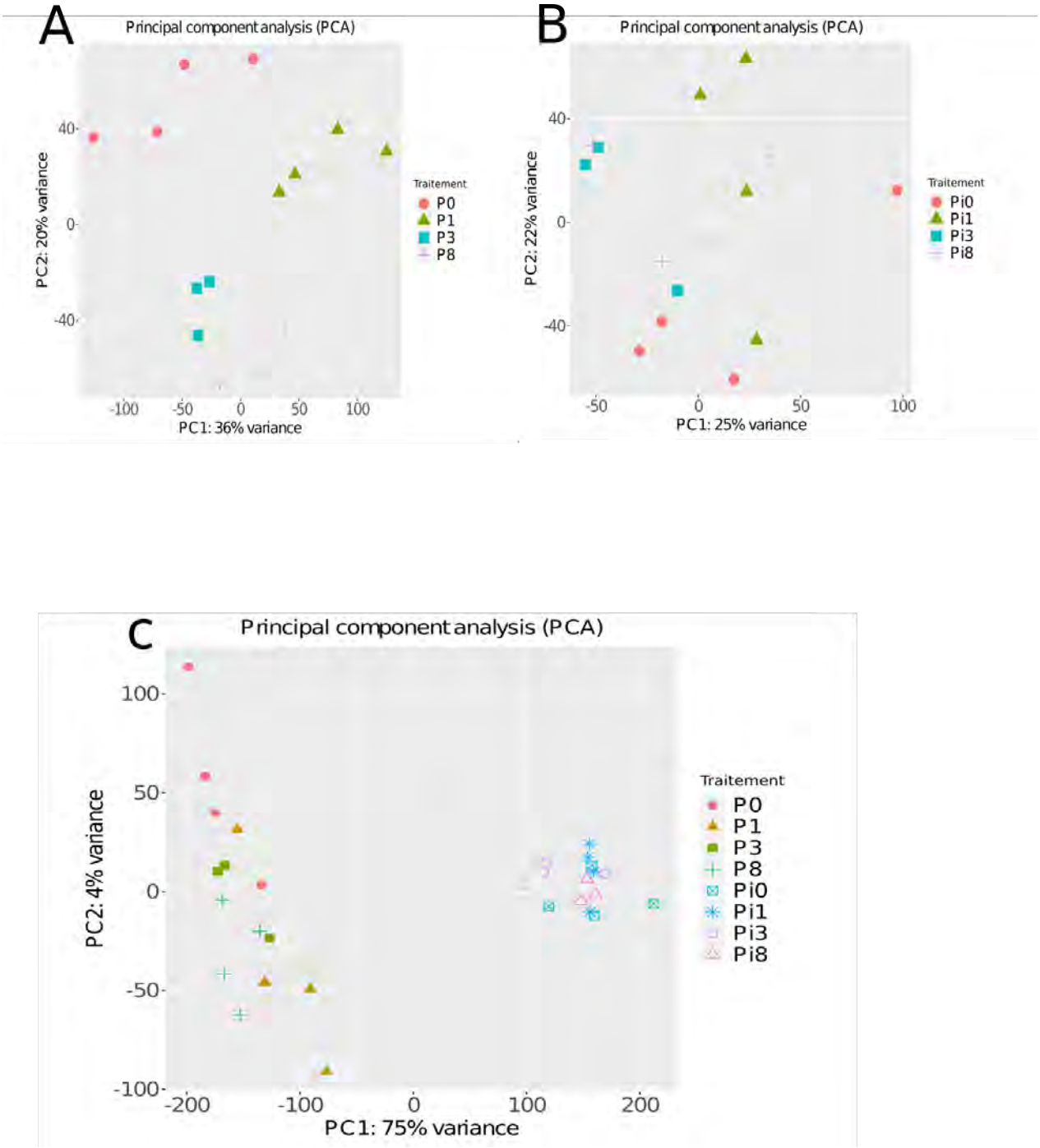


Figure 5.2: Cohérence des données RNA-seq. **A**: ACP des échantillons des plantes soumises à un nombre différent de stimulations acoustiques. **B**: ACP des plantes soumises à un nombre différent de stimulations acoustiques puis infectées. **C**: ACP sur tous les échantillons. PN correspond aux échantillons des plantes saines, PiN à ceux des plantes infectées et N est le nombre de stimulations acoustiques répétées.

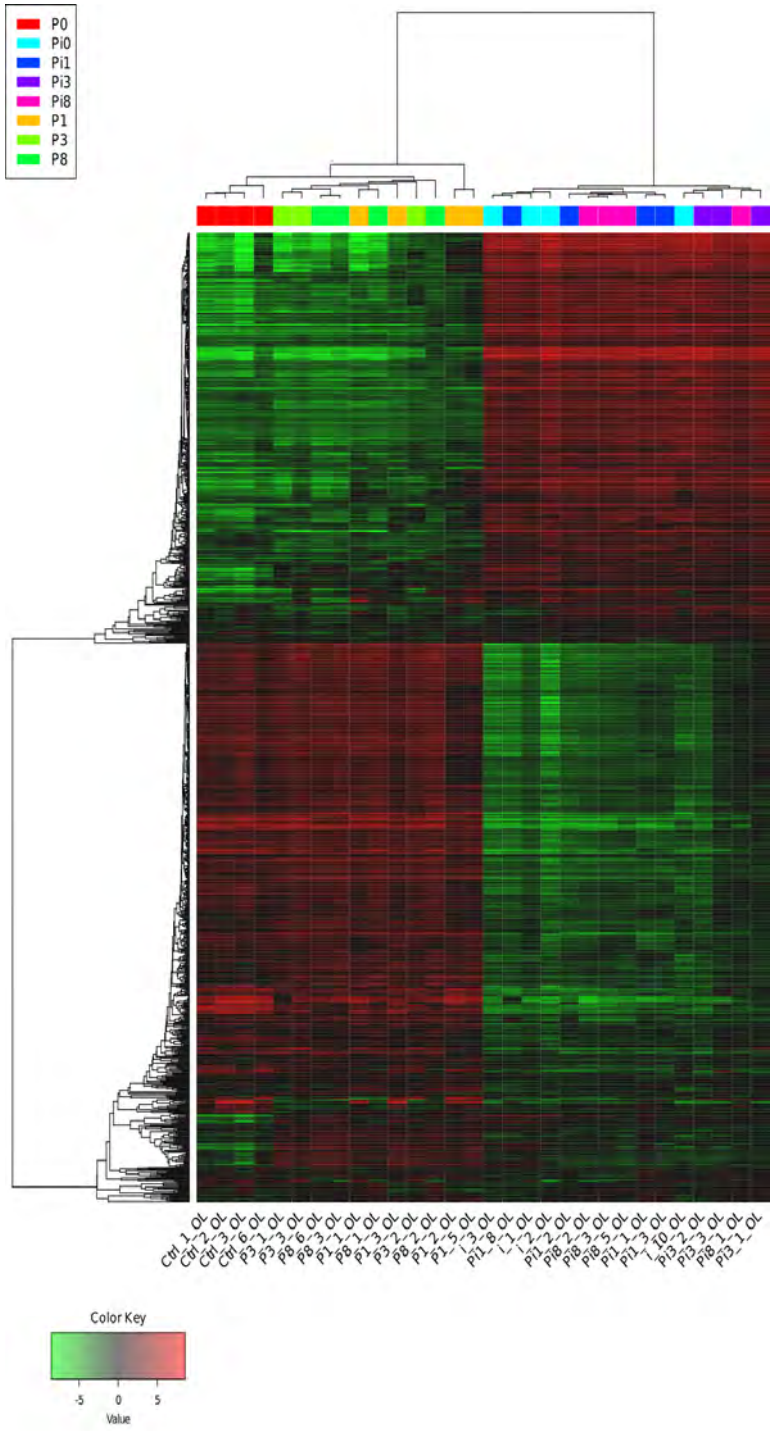


Figure 5.3: Heatmap des expressions des 1000 gènes les plus modulés classés par gènes (lignes) et par échantillons (colonnes) RNA-seq. PN correspond aux échantillons des plantes saines, PiN à ceux des plantes infectées et N est le nombre de stimulations acoustiques répétées.

6 Reprogrammation massive du transcriptome d'*Arabidopsis thaliana* en réponse au son

Sommaire

Les études déjà publiées rapportent que l'exposition des plantes à des simulations sonores répétées active de nombreux mécanismes biologiques en relation avec l'immunité. La modulation de certains gènes contribue à l'amélioration de la réponse immunitaire de la plante [Ghosh et al., 2016, 2017, Qi et al., 2010, Kwon et al., 2012]. La répétition de stimulations sonores module les concentrations en acide jasmonique et acide salycilique. Ces acides sont des hormones impliquées dans la signalisation de la réponse immunitaire. On admet souvent pour simplifier que l'augmentation de la concentration en acide salicylique est importante pour la défense contre les bactéries et que l'augmentation en acide jasmonique est important pour la défense contre les champignons nécrotrophes. Au cas par cas, ces relations semblent bien plus complexes et interdépendantes. Les descriptions déjà publiées de l'effet intégratif de simulations sonores répétées présentent un autre biais. En effet, elles ont été réalisées par analyses de puces ARN (https://www.ebi.ac.uk/arrayexpress/; accession no. E-MTAB-4077) qui

permettent d'évaluer l'expression d'une partie des gènes seulement.

Au cours de ce travail, nous voulions décrire à l'échelle du génome l'effet des répétions sonore en nous focalisant sur la compréhension de l'effet des répétitions plutôt que de l'effet intégré. Nous avons donc mesuré les modifications transcriptomiques ayant lieu dans des feuilles d'*Arabidopsis thaliana* âgées de 4 semaines après des expositions répétées au son (3h/jour avec une fréquence de 1 kHZ à 100 dB). Comme la QDR est fortement multigénique et n'est pas en aval des voies de l'acide jasmonique ou salycilique, nous avons priviligé une approche sans *a priori* à l'echelle de génome. Une partie du jeu de données permet d'évaluer l'expression des gènes après 1,3 et 8 jours d'exposition au son et donc d'explorer la mémoire transcriptionnelle. L'autre partie évalue l'expression des gènes des plantes exposées au son 1, 3 et 8 jours puis infectées et permet de mieux comprendre le priming.

Nous avons utilisé le package DESeq2 [Love et al., 2014] pour effectuer l'analyse des gènes differentiellement exprimés. Les données ont d'abord été normalisées à l'aide de la méthode de la médiane des ratios proposée par DESeq2. Nous avons considéré l'expression des gènes pour les plantes saines non exposées au son comme référence de l'analyse. Les gènes différentiellement exprimés pour les plantes exposées à 1, 3, 8 stimulations acoustiques ont été identifiés par trois tests statistiques appariés basés sur la distribution binomiale négative. Nous avons considéré comme différentiellement exprimés les gènes associées aux p-valeurs ajustées par la méthode de Bonferroni [Holm, 1979] inférieures à 5%. Aucun seuil n'a été fixé sur les log_2 *Fold change*. Cette méthode d'analyse différentielle est détaillée dans le 3.

6.1 La répétition des stimulations sonores provoque la reprogrammation massive du transcriptome

Pour mieux comprendre les bases moléculaires associées à l'augmentation progressive de la résistance des plantes primées par des stimulations acoustiques répétés, nous avons identifié les gènes différentiellement exprimées chez des plantes saines d'*Arabidopsis thaliana* Col-0 exposées à des stimulations acoustiques répétées. Les stimulations acoustiques répétées ont provoqué une reprogrammation massive du transcriptome et 35 % des gènes d'*Arabidopsis thaliana* ont été modulés au moins une fois par une stimulation. L'analyse temporelle des gènes exprimés différentiellement montre la dépendance entre le nombre de répétions sonores et le

nombre de gènes modulés (Figure 6.1.A). Entre 10 % et 24 % du génome est modulé par les stimulations acoustiques. La plus petite proportion de gènes modulés a été trouvée pour 3 stimulations acoustiques (Figure 6.1.A). Le nombre limité de gènes exprimés après 3 stimulations peut indiquer un mécanisme de dégradation de l'ARN. Seuls 1422 (15 %) des 9554 gènes modulés par des stimulations acoustiques répétées ont été exprimés indépendamment du nombre de stimulation, alors que 5578 (58 %) ont été modulés par un nombre spécifique de stimulations, ce qui montre un effet de la mémoire transcriptionnelle dominant sur l'expression des gènes.

Comme déjà rapporté, la QDR est un processus fortement multigénique qui implique un nombre considérable de gènes [Sucher et al., 2020]. Environ 50% (14040) des gènes d'*Arabidopsis thaliana* sont différentiellement exprimés en réponse à l'infection par *S. sclerotiorum* (Figure 6.1.A). L'analyse différentielle montre que les stimulations acoustiques modulent jusqu'à 33 % des gènes associés à la QDR chez des plantes saines, avec un minimum de 14 % observé après 3 stimulations (Figure 6.1.B). Dans 80% des cas, les stimulations acoustiques modulent les gènes dans le même sens que l'infection.

L'analyse d'enrichissement d'ontologie en processus biologique des gènes (GO) modulés par les stimulations acoustiques montre que la mémoire transcriptionnelle ne résulte pas seulement de l'activation de mécanismes déclenchés après la première stimulation, mais résulte de l'activation des mécanismes successifs (Figure 6.1C, D). La première stimulation acoustique conduit à l'activation de processus de réponse à différents stimuli, l'activation de réponse à une stimulation abiotique probablement en aval de la perception de la stimulation acoustique, comme la réponse au stress osmotique et au stimulus endogène (Figure 6.1C). Les premières stimulations conduisent aussi à l'activation de processus associés aux interactions plantes-pathogènes, comme la réponse à la chitine et la réponse à d'autres organismes. L'exposition à 3 stimulations conduit à la modulation d'autres processus biologiques associés aux réponses aux oxoacides communément sécrétés par les champignons nécrotrophes et à la biosynthèse d'éléments structurels glucidiques (Figure.6.1C). Enfin, après 8 stimulations, l'analyse d'ontologies révèle que la mise en place de nouveaux processus biologiques est anecdotique. L'analyse de l'enrichissement des ontologies des composants cellulaires indique que la mémoire transcriptionnelle induite par la répétions des stimulations acoustiques est composée d'acteurs clés de la QDR (Figure 6.1D).

Les analyses d'enrichissement suggèrent que la mécanoperception et l'osmoperception sont deux voies impliquées dans la mémoire transcriptionnelle activée par les stimulations répétées. Ces voies de signalisation sont également impliquées dans la régulation de la réponse immunitaire des plantes. Les canaux de type mécanosen-

sible (AtMSL2, 3, 4, 6, 9 10) et AtMCA1 (MID1-complementing activity) sont impliqués dans l'accumulation d'acide jasmonique et le dépôt de lignine [Engelsdorf et al., 2018a].

L'article de Léger et al. (2022, montre que l'immunité déclenchée par la mécano-perception (Mechano-signalling triggered immunity : MTI), basée sur la configuration anisotrope des microtubules corticaux, améliore la défense des plantes et contribue à 40% de la résistance [Léger et al., 2022]. L'analyse différentielle des expressions des gènes confirme l'implication de la mécano-perception et de l'osmo-perception dans la mémoire transcriptionnelle et la réponse aux stimulations acoustiques. Les canaux mécanosensibles AtMSL5 et AtMSL10 sont des gènes de réponse aux stimulations acoustiques et sont modulés avec la même amplitude après chaque stimulation alors que AtMSL6 est impliqué dans la mémoire transcriptionnelle et modulé par 1 et 8 stimulations. AtMSL2, AtMSL3, AtMSL4, AtMCA1 ne sont pas modulés par les stimulations acoustiques, mais modulés uniquement par l'infection. En analysant nos données transcriptomiques nous avons identifié 142 gènes associés à la MTI. Ces gènes qui sont exprimés lors de l'infection, sont également modulés par les stimulations acoustiques et impliqués dans la mémoire transcriptionnelle (Figure 6.1.E). 72 % des gènes associés à la MTI sont modulés après la première stimulation, 35 % après 3 stimulations et 60 % après 8 stimulations. 18 % des gènes associés à la MTI sont modulés indépendamment du nombre de stimulations acoustiques et 23 gènes sont des gènes de réponse aux stimulations acoustiques.

Les facteurs de transcription sont des éléments clés liés à la mémoire transcriptionnelle. 10 des 31 régulateurs épigénétiques impliqués dans la mémoire thermique et la résistance [Liu et al., 2022] sont modulés par les stimulations acoustiques chez des plantes saines (Figure 6.1 F). Pour tester le rôle des facteurs de transcription (TF) dans l'évolution de la mémoire transcriptionnelle, nous avons analysé les TF qui sont modulés par les stimulations acoustiques (Figure 6.1 F, G). L'expression des facteurs de transcription dépend du nombre de stimulations et seuls 117 (20 %) TF connus d'*A. thaliana* sont modulés indépendamment. Tous les TF associés à la MTI (18) sont modulés par les stimulations, confirmant que les voies de la MTI sont impliquées dans l'évolution dynamique de la mémoire transcriptionnelle.

Ces résultats ont montré que les stimulations acoustiques répétées déclenchent une mémoire transcriptionnelle associée à de nouveaux processus biologiques. Cette mémoire transcriptionnelle, assurée par des facteurs de transcription et impliquant les voies de la MTI et de la mécano-perception, permet la mise en place progressive de nouveaux processus biologiques associés à l'augmentation des résistances chez *A. thaliana* .

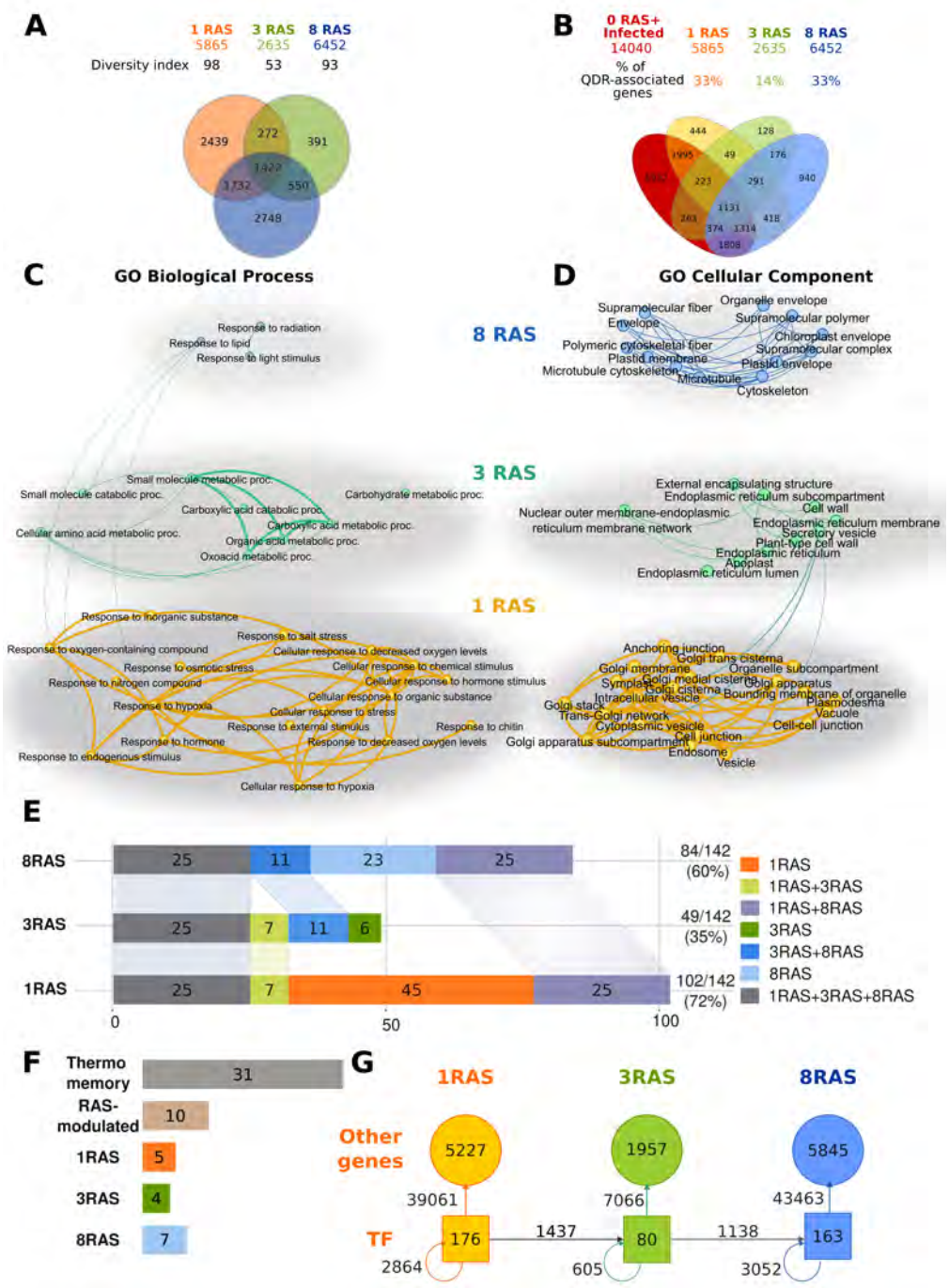


Figure 6.1: Effet des répétitions de simulation sonores sur l'expression des gènes chez les plantes saines. A. Diagramme de Venn des gènes exprimés après 1, 3, 8 stimulations dans des plantes non infectées par rapport à une plante non soumise aux stimulations acoustiques. B. Diagramme de Venn des gènes exprimés après 1, 3 et 8 stimulations dans des plantes non infectées par rapport aux plantes non soumises aux stimulations acoustiques mais infectées par *Sclerotinia sclerotiorum*. Les chiffres indiquent le nombre total de gènes différentiellement exprimés. C. Analyses d'enrichissement des ontologies de processus biologiques. D. Analyses d'enrichissement des ontologies de composants cellulaires. E. L'expression différentielle de gènes associés à l'immunité déclenchée par la mécanoperception. F. Diagramme de Venn des facteurs de transcription modulés. G. Rôle des facteurs de transcription dans la mémoire transcriptionnelle.

6.2 L'effet de répétition de stimulation sonore sur les plantes infectées

La mémoire transcriptionnelle est un mécanisme fondamental dans l'adaptation des plantes à des environnements fluctuants. Pour explorer le lien moléculaire entre la mémoire transcriptionnelle activé par des stimulations acoustiques répétées et le priming du QDR des plantes, nous avons étudié la différence entre les transcriptomes de plantes saines et les transcriptomes de plantes exposées à un nombre croissant de stimulations sonores puis infectées par *Sclerotinia sclerotiorum* . Des plants d'*A. thaliana* Col-0 âgées de 4 semaines ont été exposés au son de 0 à 8 fois (avec la même fréquence de 1 kHz pendant 3h par jour). Après la dernière stimulation acoustique, les plantes ont été inoculées avec le champignon nécrotrophe *Sclerotinia sclerotiorum* . L'extraction des ARN a été effectuée 24 heures après l'inoculation. Les gènes différentiellement exprimés sont identifiés en comparant les expressions de gènes dans les plantes exposées au son et infectées aux plantes saines (Figure 6.2).

Comme souvent observé avec les stress séquentiels, les transcriptomes sont dominés par le dernier stress subi sur la plante [Desaint et al., 2021]. 13480 (96%) des 14040 gènes associés à la QDR modulés lors de l'infection chez les plantes saines ont également été modulés dans la même direction chez les plantes primés (Figure 6.2.A). Cette proportion est restée stable pour les gènes sur et sous-exprimés, sauf une légère diminution observable après 3 stimulations acoustiques (85%) (Figure 6.2.B) .

Les plantes exposées aux stimulations sonores avant l'infection présentent des variations transcriptomiques dans environ 10 % des gènes. Pour tester si ces changements dépendent de l'interaction entre la mémoire transcriptionnelle et le contexte d'infection, nous avons comparé les transcriptomes de plantes saines exposés à des stimulations acoustiques aux transcriptomes de plantes infectées exposées et non exposées à des stimulations acoustiques (Figure 6.2.C). La plupart des gènes exprimés dans les plantes exposées aux stimulations acoustiques avant l'infection ne sont modulés que dans le contexte de l'infection. Environ 1 000 gènes ont été modulés dans des plantes stimulées avant l'infection, quel que soit le nombre de stimulations (Figure 6.2.C).

Cette analyse différentielle montre l'activation de nouveaux gènes qui peuvent être impliqués dans la QDR des plantes exposées aux stimulations acoustiques. Pour chaque durée de stimulations, environ 1000 gènes sont modulés dans les plantes stimulés uniquement dans le contexte d'une infection et ne sont pas modulés chez les plantes seulement infectées (Figure 6.2.C). 348 de ces 1000 gènes sont exprimés

indépendamment du nombre de stimulations (Figure 6.2.D).

L'analyse d'enrichissement d'ontologie des gènes (GO) montre que les processus biologiques activés dans les plantes exposées aux stimulations acoustiques répétées évoluent avec le nombre de stimulations et augmentent le nombre de processus impliqués dans la réponse de défense (Figure 6.2.E). Après 3 stimulations, les plantes infectées présentent un nombre plus élevé de gènes associés à la réponse de défense aux bactéries. Après 8 stimulations, les plantes montrent également une réponse plus élevée aux lipides et aux ROS (reactive oxygen species) (Figure 6.2.E). Après 3 stimulations, les plantes infectées présentent des différences de protéasome, pouvant expliquer le nombre inférieur de gènes modulés (Figure 6.2.F).

Les stimulations acoustiques successives ont conduit à la diversification des voies de la QDR activées chez les plantes. La mémoire transcriptionnelle est également un mécanisme essentiel du priming des gènes. Les gènes primés se caractérisent par des réponses plus rapides et plus fortes [Avramova, 2015]. Nous avons identifié des gènes primés dans l'ensemble des gènes associés à la QDR et dans l'ensemble des gènes recrutés. Comme les gènes sont modulés dans le contexte d'une infection de manière plus intensive dans les plantes exposées à des stimulations que dans les plantes naïves. 399 des gènes du priming sont modulés de manière plus intensive chez les plantes exposées à 3 et 8 stimulations par rapport aux plantes naïves et exposées à 1 stimulation. 22 gènes parmi les gènes recrutés sont des gènes putatifs de priming. Deux gènes associés au MTI sont également associés au priming ; AtCYP707A1 (AT4G19230) est impliqué dans la réponse de défense contre les champignons, tandis qu'AtGIG1 (AT3G57860) est un régulateur positif de la défense des plantes et contrôle l'entrée dans la deuxième division méiotique.

Ces résultats ont montré que la mémoire transcriptionnelle déclenchée par des stimulations acoustiques répétées conduisait à une amélioration de la réponse immunitaire par deux mécanismes : la diversification de la réponse par le recrutement de gènes et le priming des gènes de QDR.

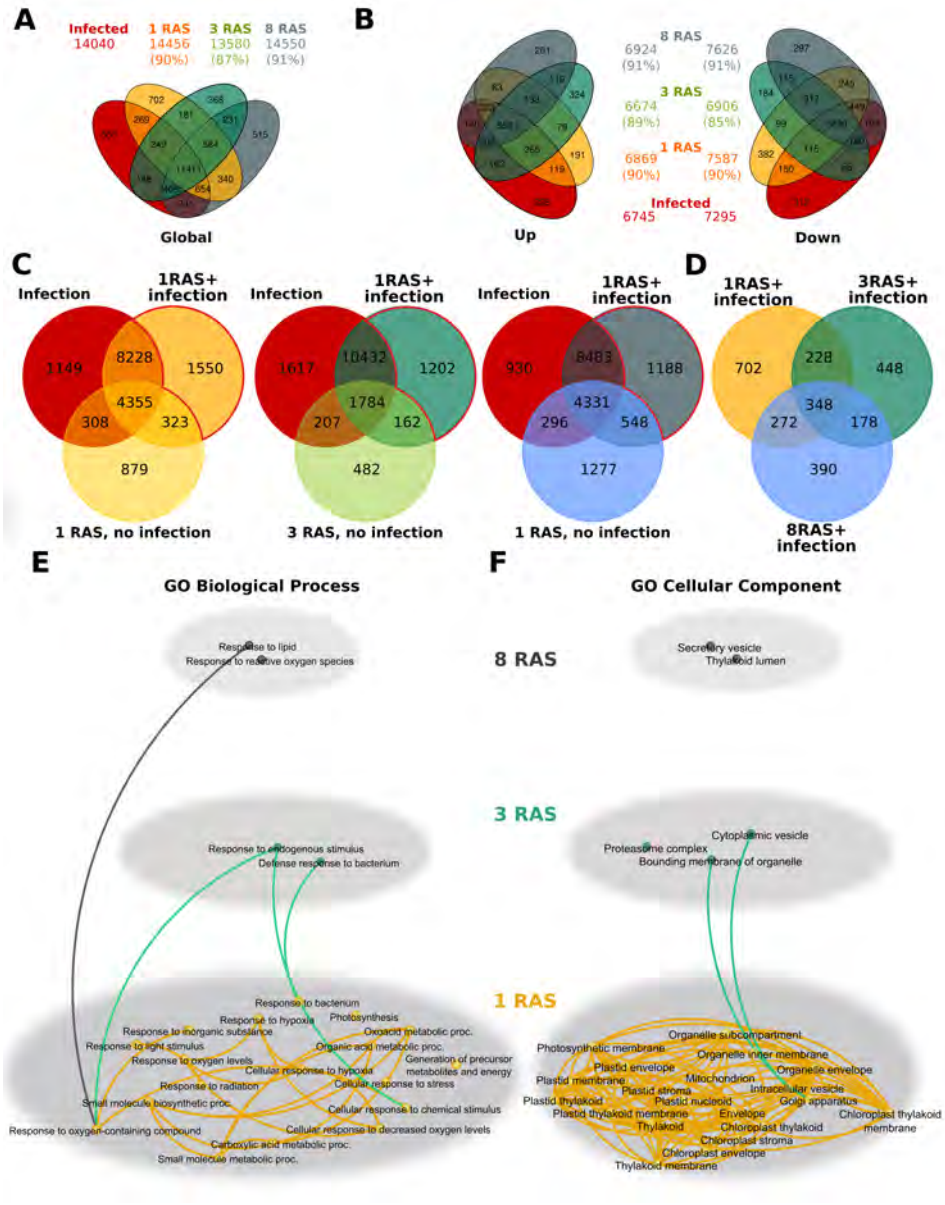


Figure 6.2: Expression des gènes chez les plantes exposées aux simulations sonores puis infectées. A. Diagramme de Venn des gènes exprimés après 1, 3, 8 simulations sonores dans une plante infectée par rapport à une plante naïve. B. Diagrammes de Venn pour les gènes sur et sous-exprimés. C. Diagrammes de Venn des gènes exprimés lors de l'infection, après des simulations sonores chez des plantes saines et infectées. D. Diagramme de Venn des gènes exprimés dans les plantes exposées aux stimulations sonores avant l'infection et modulé uniquement dans le contexte de l'infection. E. Analyses d'enrichissement des ontologies de processus biologiques. F. Analyses d'enrichissement des ontologies de composants cellulaires.

6.3 Discussion

L'analyse différentielle montre que la mémoire transcriptionnelle associée aux stimulations acoustiques répétées déclenche l'expression de 102 gènes associés à la MTI (Mechano-Signaling Triggered Immunity) chez des plantes saines [Léger et al., 2022]. Ces gènes sont liés à la régulation de la défense des plantes par la perception de signaux mécaniques crée par des agents pathogènes.

Les régulations épigénétiques jouent un rôle important dans les processus d'acclimatation des plantes à des signaux mécaniques répétés. Par exemple, les marques d'histones telles que H3K9/14ac et H3K4me3 régulent la désensibilisation du peuplier aux flexions répétées [Ghosh et al., 2023a]. Nos données montrent que 30 % des marqueurs de mémoire thermique et de défense sont modulés par des stimulations acoustiques répétées. Les marques épigénétiques sont des mécanismes associés à la mémoire transcriptionnelle des stress environnementaux et antagonistes de la dégradation de l'ARN [Crisp et al., 2017]. La dégradation de l'ARN est souvent liée à l'établissement d'un nouveau régime biologique fonctionnel. Nos données transcriptomiques révèlent que l'efficacité du priming survenant après 3 répétitions acoustiques s'accompagne d'une forte dégradation de l'ARN. Or, nos expériences montrent que la mémoire transcriptionnelle répétitions acoustiques persiste simultanément avec la dégradation (decay) de l'ARN. Ainsi, la dégradation de l'ARN et les modifications épigénétiques peuvent agir ensemble pour maintenir la mémoire des stress environnementaux et assurer la transition de la plante d'un état naïf à un état primé.

7 Choix des méthodes utilisées

Sommaire

L'objectif de ce chapitre est de présenter l'ensemble des méthodes sélectionnées, adaptées, développées au cours de la thèse en vue d'inférer un réseau dynamique

permettant de mieux comprendre la réponse d'*A. thaliana* au son. Les méthodes d'inférence dynamique permettent de décrire la dynamique de réseaux et de reconstituer les relations entre les gènes (ou classes des gènes), c'est pour cela qu'elles nous ont semblé les plus adéquates pour répondre à notre question biologique.

Nos données disponibles, avec un nombre de gènes modulés dans la réponse très largement supérieur au nombre de mesures (9554 gènes DEG pour 4 mesures temporelles), et des mesures temporelles espacées irrégulièrement (0j, 1j, 3j et 8j) nous ont conduit à devoir combiner classification, reconstruction de données et inférence dynamique. Ce chapitre présente les différentes méthodes qui ont été considérées.

7.1 Notations

Les notations utilisées dans la suite de ce chapitre sont définies ici. Le tableau de comptage d'expression de gènes est représenté sous la forme d'une matrice, notée X. Cette matrice est de taille $N \times M$, où M est le nombre des mesures temporelles et N le nombre de gènes. La notation X_{g,t_m} correspond à l'expression moyenne des expressions des réplicats biologiques transformées en $log(CPM)$ du gène g ($1 \leq g \leq N$) à l'instant t_m ($1 \leq m \leq M$). Pour nos données expérimentales, m varie donc de 1 à 4 avec t_m valant respectivement 0, 1, 3 et 8 jours. La notation X_t sera également utilisée, pour représenter le vecteur d'expression des gènes à la date t. Dans le cas d'utilisation de K clusters de gènes, la notation $X_{k,t}$, $k = 1, \cdots K$ sera aussi utilisée pour représenter l'expression du cluster k à la date t. Dans ce cas, X_t pourra être un vecteur de dimension K. On notera N_k la taille du cluster k.

7.2 Méthode des signatures pour la classification de trajectoires d'expression

Pour inférer un réseau dynamique de régulations des gènes, nous avons dû résoudre une première difficulté liée au très grand nombre de gènes à prendre en compte. Nous avons classiquement fait le choix de travailler avec des clusters de gènes, obtenus par classification des données d'expression. Les différentes méthodes présentées en 4.3 sont bien adaptées lorsqu'on utilise des séries temporelles assez longues et des mesures régulièrement espacées. Nous avons pu le vérifier en testant expérimentalement une méthode de classification hiérarchique temporelle. L'idée

était de découper l'intervalle en plusieurs sous intervalles pour capter le plus possible la variation dans ces séries et de prendre en compte ces variations lors de l'étape de classification. Une description détaillée de cette approche est présentée dans l'annexe 2. Néanmoins, ce n'est pas toujours le cas lorsqu'on travaille avec des séries temporelles courtes issues d'expérience biologiques, où la contrainte de coût est importante.

Nous avons donc développé une méthode assez simple de classification de séries temporelles courtes (typiquement moins d'une dizaine de mesures temporelles), aux points non nécessairement régulièrement espacés, qui repose sur un principe de regroupement basé sur la forme, la "signature", des trajectoires. Une comparaison avec des méthodes classiques de classification sera effectuée pour valider expérimentalement cette méthode.

7.2.1 Utilisation des $2-$signatures

Une première approche peut-être simplement définie sur la base des deux signes $+1$ (*croissant*) et -1 (*décroissant*). Pour chaque gène i, nous définissons le vecteur de $2-$signature $s^i = (sign(X_{i,t_{m+1}} - X_{i,t_m}),\ m = 1, ..., M-1)$ ($sign(0)$ est arbitrairement fixé à un des deux signes). Les gènes $\{1, \cdots, N\}$ de même signature sont alors regroupés dans la même classe. On obtient alors 2^{M-1} classes (ou moins s'il y a des classes vides). Si on travaille avec 4 points temporels, on aura donc $2^3 = 8$ classes.

Pour améliorer et minimiser la variance intra-classe, une classification k-means peut être appliquée ensuite pour chaque classe k ($1 \leq k \leq 2^{M-1}$). Le nombre de sous-classes c_k est déterminé en optimisant pour chaque classe le coefficient de silhouette (Rousseeuw,1987). Cette étape permet de former in fine K classes $C_k,\ k \in \{1, \ldots, K\}$, $K = \sum_{k=1}^{2^{M-1}} c_k$.

Une étude comparative est effectuée avec cette méthode de classification, pour étudier l'efficacité de cette méthode par rapport aux autres méthodes existantes lorsqu'on travaille avec des données temporelles très courtes.

7.2.2 Utilisation des $3-$signatures

La méthode des $2-$signatures travaille avec seulement deux signes $+1$ ou -1 (l'expression croit ou décroît entre deux instants), sans prendre en compte le fait que le gène soit différentiellement exprimé ou pas entre ces deux instants de mesures. Il est possible d'étendre cette méthode en travaillant avec trois signes, $+1$, -1 ou 0, en réservant les signes $+1$ et -1 entre deux instants aux seuls gènes différentiellement exprimés entre ces deux instants ($+1$ si croissance, -1 si décroissance). Pour les gènes non différentiellement exprimés entre ces deux

instants, le signe 0 est attribué (Fig. 7.1). On obtient donc ici à ce stade jusqu'à 3^{M-1} classes, soit $3^3 = 27$ clusters pour nos données.

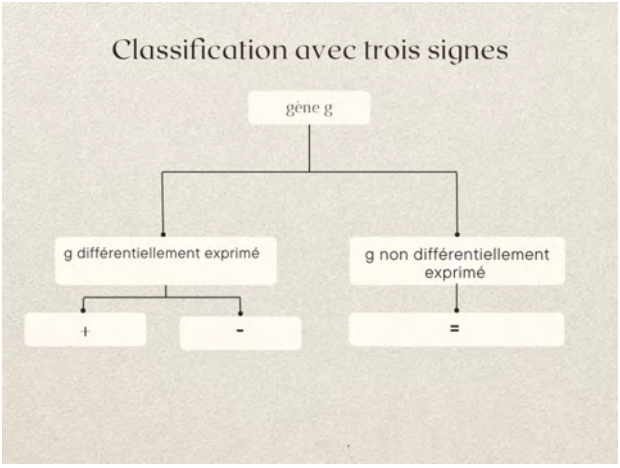


Figure 7.1: Organigramme de la classification avec les 3−signatures. Entre deux instants successifs de mesure, si un gène g est différentiellement exprimé un signe $+1$ ou -1 lui sera attribué. Si le gène n'est pas différentiellement exprimé entre ces deux instants, le signe 0 lui sera attribué.

La suite de la méthode concernant l'utilisation ou non d'une sous classification à base de k-means est similaire à celle avec les 2−signatures. Nous verrons dans le chapitre suivant un exemple de mise en oeuvre de cette méthode.

7.3 Modélisation de la dynamique d'expression

Les approches d'inférence de réseau dynamique reposent nécessairement sur un choix de modèle dynamique des expressions en chaque noeud du réseau. Les approches basées sur l'utilisation de clusters considèrent généralement comme représentant de chaque cluster sa moyenne en chaque instant, et le modèle dynamique décrit donc l'évolution de ces moyennes au cours du temps. Nous avons commencé avec cette approche, mais la dynamique de la variance intra-classe était de la sorte assez mal représentée, et nous avons choisi de développer une modélisation permettant de mieux gérer les valeurs extrêmes de chaque cluster. Une autre nécessité a été de modéliser explicitement l'effet de la stimulation sonore, en le distinguant de ce qui pourrait être vu comme la régulation dynamique naturelle du réseau.

7.3.1 Modèle autorégressif simple sur les moyennes des classes

Le premier modèle de dynamique que nous avons considéré est un modèle classique autorégressif linéaire. Dans ce modèle, l'expression moyenne des gènes d'un cluster à un instant est égale à la somme pondérée des expressions moyennes de tous les clusters à l'instant précédent. Les coefficients de pondération reflètent ainsi l'influence de l'expression des gènes des différents clusters sur l'expression des gènes du cluster prédit.

$$\bar{X}_{i,t} = \sum_{j=1}^{K} a_{i,j}\ \bar{X}_{j,t-1} + \epsilon_i \tag{7.1}$$

avec i et j deux classes parmi les K classes du réseau, et $\bar{X}_i^t$ l'expression moyenne du cluster i au temps t

$$\bar{X}_{i,t} = \frac{1}{N_i} \sum_{g \in C_i} X_{g,t}. \tag{7.2}$$

Le poids $a_{i,j}$ représente l'impact des gènes du cluster j sur l'expression moyenne des gènes du cluster i. Cette équation peut s'écrire également

$$\bar{X}_t = A\bar{X}_{t-1} + \epsilon \tag{7.3}$$

avec $A = (a_{i,j})_{i,j=1,\cdots,K}$.

7.3.2 Modélisation de l'effet direct du son

L'utilisation du modèle autorégressif précédent ne permet pas de faire une distinction claire entre l'effet direct du traitement, qui est l'exposition au son dans notre étude, et la régulation dynamique du réseau.

Nous avons donc choisi de reformuler ce modèle sous la forme d'un modèle d'équation aux différences explicite, que nous avons étendu en considérant un vecteur d'interception $b = (b_i)_{i=1,K}$ de dimension $K \times 1$, supposé indépendant du temps, et qui va représenter l'effet direct de la stimulation sonore à la date t sur la dynamique d'expression des gènes. Avec les notations précédentes, on a donc :

$$X_{t+1} - X_t = AX_t + b + \epsilon_t \tag{7.4}$$

On note ici la définition légèrement différente de A, correspondant à $I + A$ de l'équation 7.1.

Nous avons ensuite ajouté la contrainte supplémentaire $AX_0 = 0$, qui décrit le fait que l'effet du son commence seulement après un jour de son, à $t_2 = 1$, et que les gènes du système sont dans un état stable avant le début de l'expérience, à $t_1 = 0$ (Figure 7.2). Au final, le modèle a la forme suivante :

$$\begin{aligned} X_{t+1} - X_t &= AX_t + b + \epsilon_t, \quad \forall t = 0, \cdots T-1 \\ 0 &= AX_0, \end{aligned} \tag{7.5}$$

avec X_t le vecteur de dimension $K \times 1$ représentant les classes à l'instant t.

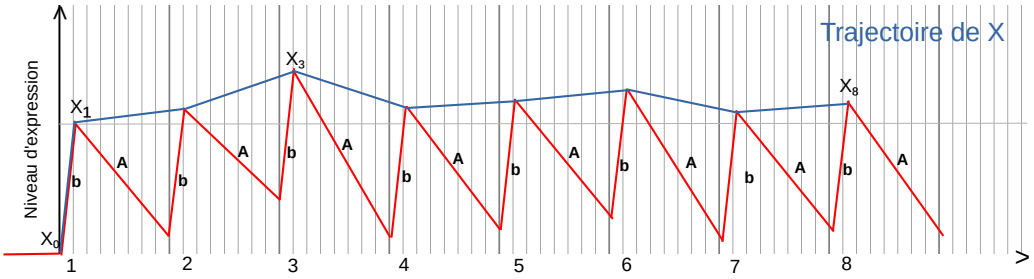


Figure 7.2: Exemple de trajectoire correspondant au modèle 7.5.

Appliqué à l'utilisation de la moyenne pour représenter l'expression des clusters, on obtient les équations suivantes :

$$\begin{aligned} \bar{X}_{t+1} - \bar{X}_t &= A\bar{X}_t + b + \epsilon_t, \quad \forall t = 0, \cdots T-1 \\ 0 &= A\bar{X}_0, \end{aligned} \tag{7.6}$$

avec $\bar{X}_t$ le vecteur d'expression moyenne des classe à l'instant t.

Avec ce modèle, nous réussissons à expliquer l'effet du son sur le système de la plante, avec A décrivant la dynamique de retour à l'équilibre des expressions et b l'effet direct de la stimulation. Néanmoins, ce modèle ne prend pas en compte explicitement la dynamique de la variabilité intra-classe et l'évolution temporelle des expressions dans chaque classe. Nous avons ainsi proposé et comparé trois familles de modèles, qui diffèrent dans leur définition de X_t, et qui conduisent à des estimations de la matrice A potentiellement différentes.

7.3.3 Prise en compte des quantiles extrêmes

Nous avons remarqué que dans chaque classe les trajectoires sont fortement parallèles entre elles (Figure 9.1.A dans le chapitre 9) , et de ce fait, pour mieux prendre en compte la variabilité de chaque cluster, nous proposons de supposer que le modèle 7.5 s'applique également aux quantiles extrêmes de la distribution de chaque cluster :

$$\begin{aligned} \bar{X}_{t+1} - \bar{X}_t &= A\bar{X}_t + b + \epsilon_t, \quad \forall t = 0, \cdots T-1 \\ X^{\alpha}_{t+1} - X^{\alpha}_t &= AX^{\alpha}_t + b + \epsilon_t, \\ X^{1-\alpha}_{t+1} - X^{1-\alpha}_t &= AX^{1-\alpha}_t + b + \epsilon_t, \\ 0 = A\bar{X}_0 &= AX^{\alpha}_0 = AX^{1-\alpha}_0, \end{aligned} \tag{7.7}$$

Dans ce modèle X^{α}_t (symétriquement pour $X^{1-\alpha}_t$) est défini comme $X_t = X^{\alpha}_{k,t k=1:K}$ et $X^{\alpha}_{k,t}$ est le quantile d'ordre α de la distribution des expressions des gènes $X_{k,t}$ au temps t. On choisira α petit (nous prendrons $\alpha = 0.05$ dans les expérimentations).

7.3.4 Utilisation de l'expression de chaque gène au sein des clusters

Toujours pour espérer mieux prendre en compte la variabilité intra-cluster, nous avons également considéré la possibilité d'appliquer ces équations dynamiques à l'échelle des gènes. Nous proposons donc un premier modèle, généralisation de l'équation 7.5, qui relie à l'échelle de chaque gène le delta d'expression avec l'expression courante du gène (à la place de la valeur moyenne du cluster) et les valeurs moyennes des autres clusters :

$$\begin{aligned} X_{g,t+1} - X_{g,t} &= A_{i,i}X_{g,t} + \sum_{j \neq i} A_{i,j}\bar{X}_{j,t} + b_i + \epsilon_t, \\ X_{g,1} - X_{g,0} &= b_i + \epsilon_0, \quad \forall t = 0, \cdots T-1, \quad i = 1, \cdots K, \quad g \in C_i. \end{aligned} \tag{7.8}$$

Une conséquence pratique de cette modélisation est qu'elle augmente considérablement le nombre d'observations prises en compte dans l'estimation de la matrice A.

7.3.5 Dépendance d'ordre entre gènes

En cherchant à trouver une possible base biologique au modèle 7.7 intégrant des contraintes sur les quantiles, nous avons imaginé une synthèse des modèles 7.7 et 7.8 en supposant une relation univoque entre les gènes dont les expressions présentent les mêmes statistiques d'ordre au sein des différents clusters, soit, dit

autrement, en supposant que les gènes qui ont une expression forte (faible) dans une classe seront influencés par les gènes qui ont des expressions fortes (faible) dans les autres classes. On aboutit alors au modèle suivant :

$$\begin{aligned} X_{g,t+1} - X_{g,t} &= A_{i,i}X_{g,t} + \sum_{j\neq i} A_{i,j}X_{j,t}^{\alpha_{g,t}} + b_i + \epsilon_t, \quad \forall t = 0, \cdots T-1, \\ X_{g,1} - X_{g,0} &= b_i + \epsilon_0, \quad \forall t = 0, \cdots T-1, \quad i = 1, \cdots K, \quad g \in C_i, \end{aligned} \tag{7.9}$$

avec dans cette formule $\alpha_{g,t} = Pr[X_{g',t} \leq X_{g,t},\ g' \in C_j]$ l'ordre du quantile $X_{g,t}$.

Il est à noter que si les distributions d'expressions de gènes au sein des classes suivent une lois normale $X_{g,t} \sim \mathcal{N}(\bar{X}_t^k, \sigma_t^k)$, ce modèle devient:

$$X_{g,t+1} - X_{g,t} = \theta_{i,t}(X_{g,t} - \bar{X}_t^i) + \sum_j A_{i,j}\bar{X}_t^j + b_i + \epsilon_t, \tag{7.10}$$

avec $\theta_{i,t} = \sum_j \frac{\sigma_t^j}{\sigma_t^i} A_{i,j}$, qui est à rapprocher du modèle 7.8 où le $\theta_{i,t}$ correspondant vaut $A_{i,i}$.

7.4 Inférence du réseau dynamique

7.4.1 Reconstruction des données manquantes

Sur la base d'un modèle dynamique supposé et d'observations temporelles des valeurs d'expression, les méthodes d'inférence vues au chapitre 4 permettent de reconstruire un réseau possible expliquant les données. Les pas de temps entre les mesures d'expressions doivent toutefois être constants pour inférer un réseau qui a du sens, et c'est pourquoi nous avons choisi d'ajouter des points de mesures artificiels en utilisant une méthode d'interpolation, qui peut être une méthode de type LOESS (LOcally Estimated Scatterplot Smoothing), ou également une simple interpolation linéaire :

$$\begin{aligned} X_{i,t} &= X_{i,m} \text{ si } t = t_m \\ &= X_{i,m} + \frac{t - t_m}{t_{m+1} - t_m}(X_{i,m+1} - X_{i,m}) \text{ si } t_m < t < t_{m+1} \end{aligned} \tag{7.11}$$

Nous avons utilisé la méthode LOESS pour inférer un premier réseau en utilisant le modèle 7.1. Pour les autres modèles, nous avons préféré utiliser l'interpolation linéaire, qui nous semble ajouter moins d'information a priori. L'étude détaillée de l'inférence des réseaux en utilisant ces deux méthodes est présentée dans le chapitre

suivant.

7.4.2 Inférence du réseau

Les différents modèles que nous avons considérés peuvent être transformés en un ensemble des modèles linéaires de régression classique pour chaque classe $k = 1, \cdots K$, sous la forme :

$$Z^k = Y(A^k)' + b^k + \epsilon_k \tag{7.12}$$

avec A^k la ligne k de la matrice A, b^k le k-éme élément de vecteur b qui décrit l'effet de perturbation extérieur sur la classe k, Y la matrice de prédiction et Z^k le vecteur réponse.

Cela signifie que pour chaque modèle de régression associé à chaque classe k nous allons déterminer les coefficients associées (les coefficients de la ligne A_k et b_k). Les dimensions exactes de Z^k et Y dépendent du modèle utilisé :

- Z^k est de dimension T, et Y est de dimension $T \times K$ pour le modèle 7.1.
- Z^k est de dimension T, et Y est de dimension $T \times K$ pour le modèle 7.6.
- Z^k est de dimension $3 * T$, et Y est de dimension $(3 * T) \times K$ pour le modèle 7.7.
- Z^k est de dimension $N_k * T$, et Y est de dimension $(N_k * T) \times K$ pour le modèle 7.8.
- Z^k est de dimension $N_k * T$, et Y est de dimension $(N_k * T) \times K$ pour le modèle 15.1.

Pour le modèle 7.1, le package R SIMONE (Statistical Inference for MOdular Network) permet d'inférer un réseau à partir des données d'expression. SIMONE génère différents réseaux en fonction de la valeur du paramètre de pénalité. Les réseaux inférés peuvent être sélectionnés soit sur sur la base du critère AIC (critère d'information d'Akaike) ou BIC (critère d'information bayésien), soit en fournissant explicitement le nombre d'arêtes. L'utilisation de ce modèle est présenté dans le chapitre suivant, ainsi qu'une étude des résultats obtenus par ce modèle.

Pour les autres modèles, nous avons fait le choix d'inférer une matrice creuse de coefficients A et d'estimer le vecteur b en utilisant la méthode d'estimation LASSO [Tibshirani, 1996].

Cette méthode est implémenté dans le package R **glmnet** [Hastie and Qian, 2014].

Une étude comparative de ces différents modèles en termes d'erreur de prédiction est présentée dans le chapitre suivant. Cette étude nous a permis de valider l'intérêt

d'introduire une prise en compte explicite de l'effet du son via le terme b, et de sélectionner le meilleur modèle possible pour l'inférence de réseau de la mémoire transcriptionnelle associées aux stimulations acoustiques répétées.

8 Tests et validations des méthodes utilisées

Sommaire

Ce chapitre présente plusieurs parties qui servent principalement à valider expérimentalement des méthodes utilisées pour inférer le réseau de réponse de la plante aux stimulations acoustiques répétées.
On va commencer par l'inférence de réseau sur nos données réelles en utilisant deux différents modèles (avec et sans terme constant). Le but est de choisir le meilleur modèle pour décrire nos données en termes d'erreur de prédiction. La deuxième étape consistera d'étudier la performance de ces deux modèles en utilisant des données simulées, en variant le nombre de noeuds et de mesures temporelles.

La première partie consiste à l'application d'un modèle autorégressif classique aux données son. La deuxième partie consiste à une étude comparative pour valider le choix de représentant des classes. La troisième partie, consiste à une étude par des données simulées pour valider le modèle qu'on va l'utiliser pour l'interprétation biologique de l'effet du son sur les plantes. La dernière partie consiste à une étude comparative pour valider notre méthode de classification en utilisant des données réelles et différentes méthodes de classification classiques.

8.1 Modèle autorégressif simple sur les moyennes des classes

Le premier modèle considéré ici est le modèle autorégressif 7.1 présenté dans le chapitre précédent. L'objectif principal ici est d'étudier l'impact sur la performance du modèle de l'absence du terme constant décrivant l'effet direct du son. On va étudier la performance en utilisant l'erreur de prédiction associée à la prédiction de chaque moyenne de classe.

Le pas de temps entre les mesures d'expression est implicitement supposé constant dans le modèle. Nous avons donc ajouté des points de mesures en utilisant ici la méthode d'interpolation LOESS (LOcally Estimated Scatterplot Smoothing), pour obtenir au total 17 points de mesure (un point toutes les 12 heures).

Pour illustrer l'inférence de réseau, nous avons utilisé l'algorithme implémenté dans le package SIMONE pour inférer le réseau de réponse de la plante au son. Dans cette partie, nous avons considéré le réseau minimisant le critère BIC (Bayesian Information Criterion).

Réseau obtenu

Pour inférer le réseau, nous avons considéré l'expression moyenne de 40 classes et les 17 points de mesures interpolés. Les classe ont été obtenues en utilisant la méthode de classification à base de de 2-signatures pour obtenir dans un premier temps 8 classes, puis 40 classes après une division des classes en 5 sous-classes (nombre de sous classes déterminé en utilisant l'indice de silhouette). Dans le réseau résultant, toutes les classes ne sont pas reliées entre elles. Nous pouvons distinguer deux catégories de noeuds, des noeuds entrants qui ne sont pas régulés par d'autre noeuds et des nœuds fils qui eux peuvent réguler d'autres noeuds (Figure 8.1).

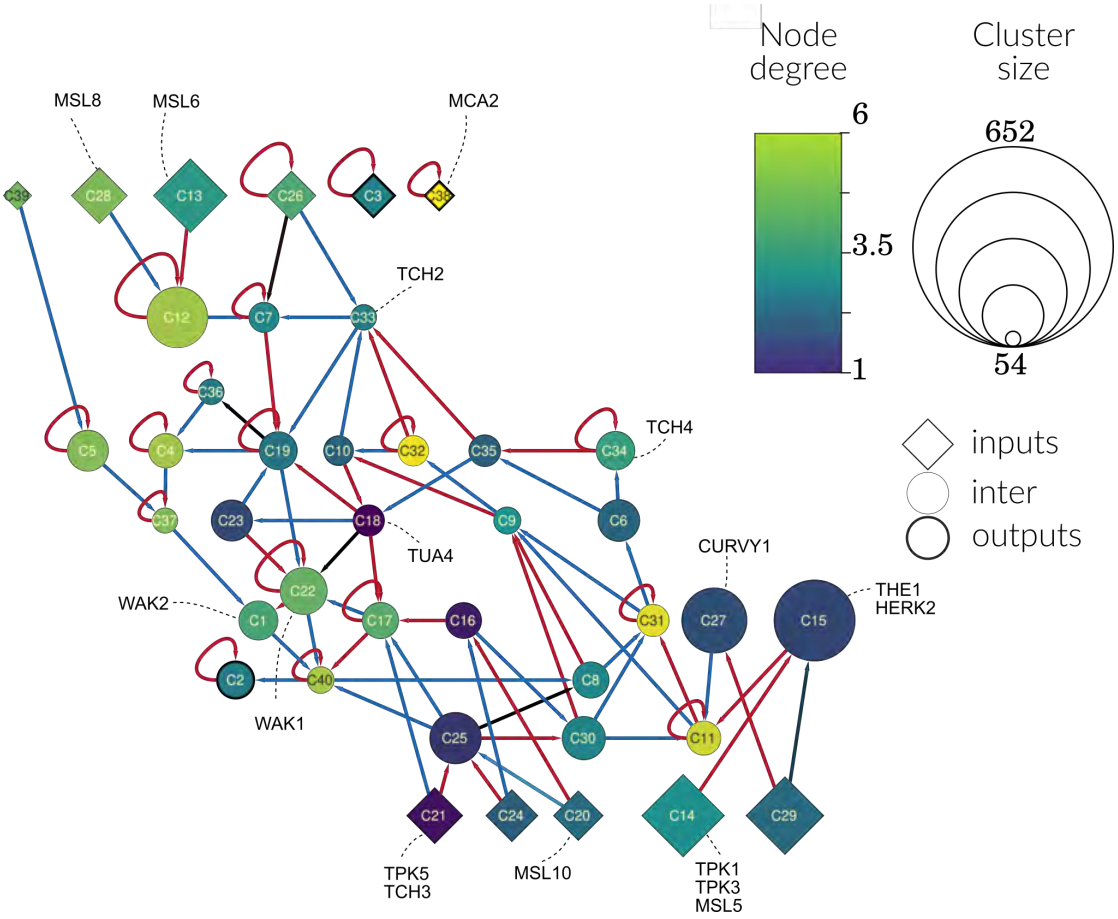


Figure 8.1: Réseau de gènes obtenu avec un modèle autorégressif simple, à partir de l'expression moyennes de 40 classes.

Erreur de prédiction

Pour évaluer la pertinence de ce modèle nous avons décidé de calculer l'erreur de prédiction de l'expression moyenne des gènes de chaque classe prédite par le modèle 7.1. L'expression moyenne prédite d'une classe est notée $\hat{X}$. Pour chaque classe i

et instant t, la valeur moyenne prédite $\hat{X_{i,t}}$ vérifie alors :

$$\hat{X_{i,t}} = \sum_{j=1}^{K} a_{i,j} \; X_{j,t-1} \tag{8.1}$$

avec $X_{j,t-1}$ l'expression moyenne prédite ou mesurée de la classe j au temps $t-1$, et $a_{i,j}$ le coefficient de la matrice A représentant l'effet de la classe j sur la classe i.

Nous avons ensuite calculé l'erreur quadratique moyenne normalisée de prédiction (NMSE) associée à ce réseau défini par sa matrice A.

Nous avons choisi de définir cette erreur selon :

$$err = \sum_{m=2}^{M} \|X_{t_m} - \tilde{A}\hat{X}_{t_m-1}\|^2 / \sum_{m=2}^{M} \|X_{t_m}\|^2 \tag{8.2}$$

où X_{t_m} représente le vecteur d'expression des N gènes observé à l'instant t_m, $\hat{X}_{t_m-1}$ le vecteur des expressions moyennes prédites ou observées à l'instant précédent $t_m - 1$, et où $\tilde{A}$ est définie comme une matrice par bloc de dimension $N \times N$ telle que $\tilde{a}_{i,j} = \frac{a_{k,l}}{N_l}$ pour un gène i dans le cluster C_k et le gène j dans le cluster C_l. L'idée ici est d'intégrer dans l'erreur de prédiction l'effet du clustering. Puisque la structure de $\tilde{A}$ est spéciale (par bloc), nous pouvons reformuler l'expression de l'erreur comme :

$$err = 1 - \frac{2Tr(CAV) - Tr(CASA')}{Tr(W)} \tag{8.3}$$

où C est la matrice diagonale $C_{kk} = N_k$, $k = 1, K$, et V, W et S sont les matrice de taille $K \times K$ de la variance-covariance empirique $V = \bar{X}_{-M}\bar{X}'_{-1}$, $W = \bar{X}_{-1}\bar{X}'_{-1}$ et $S = \bar{X}_{-M}\bar{X}'_{-M}$, où $\bar{X}_{-l}$ désigne la matrice de données moyennées $\bar{X}$ privée de sa colonne l.

L'erreur de prédiction moyenne obtenue ainsi est de 0.893. Ce résultat n'apparaît pas comme très bon, comme l'illustre la figure 8.3, car il signifie que le modèle améliore à peine la variance intrinsèque des données ($err = 1$). Nous avons aussi remarqué une hétérogénéité entre les erreurs de prédictions des moyennes des classes (Figure 8.2).

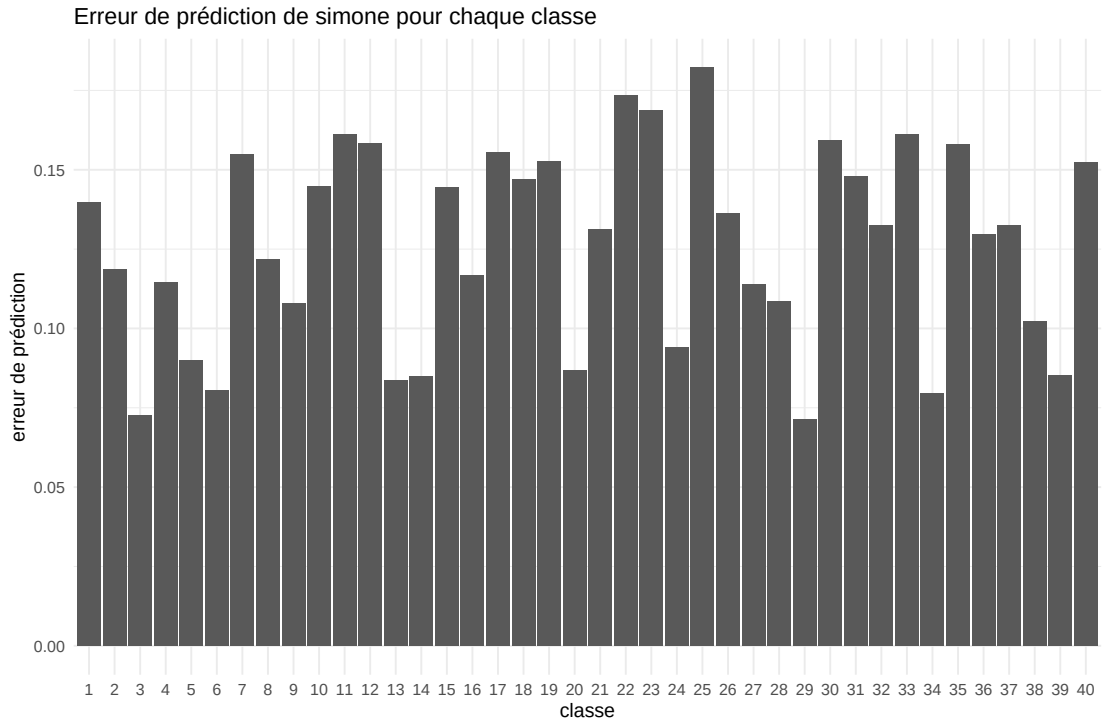


Figure 8.2: Erreur de prédiction de l'expression moyenne de chaque classe avec un modèle autorégressif simple. L'erreur calculé ici correspond à la moyenne des valeurs absolues des différences entre les expressions mesurées et prédites de chaque moyenne de chaque classe.

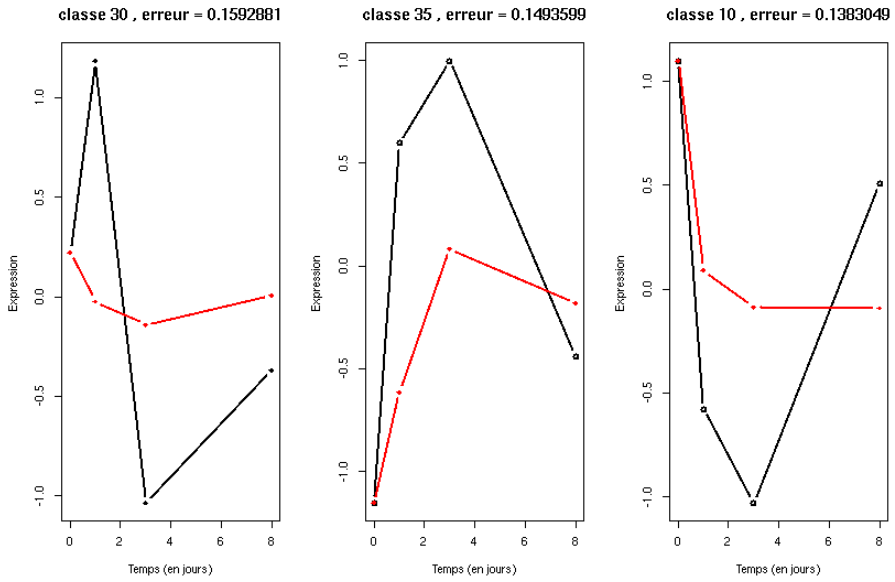


Figure 8.3: Comparaison entre l'expression moyenne (noir) et l'expression prédite (rouge) pour trois différentes classes parmi les 40 classes.

En conclusion, l'utilisation d'un modèle autorégressif classique sur les expressions moyenne de chaque classe ne permet pas vraiment d'inférer un réseau satisfaisant, du fait du peu de confiance dans la prédiction associée. Nous pouvons expliquer ce problème de modélisation par le fait que le nombre de classes est trop important en comparaison du nombre de mesures, mais aussi par la forme de l'équation linéaire, non différentielle, basée sur la seule moyenne des classes, et sans effet constant du son.

Nous avons donc essayé de travailler avec les différents modèles présentés au chapitre précédent, pour voir en particulier s'ils pouvaient améliorer l'erreur de prédiction.

8.2 Évaluation de différents modèles d'inférence

Pour comparer entre les différents modèles présentés dans le chapitre 7, nous avons utilisé les 8 classes obtenus en utilisant la méthode de classification à 2 signatures (validation de la méthode dans le section 8.4) et nous avons interpolé linéairement nos 4 points de mesures pour au final, obtenir 8 points de mesures (1 point par jour). Selon le modèle, nous avons utilisé les différentes caractéristiques des classes. Nous avons ensuite calculé l'erreur quadratique moyenne normalisée de prédiction (NMSE) associée à chaque matrice A estimée de chaque modèle. Dans le même esprit que précédemment (Eq. 15.4), nous avons défini cette erreur sur les expressions des gènes :

$$err = \sum_{g=1}^{N} \sum_{m=2}^{M} \|X_{g,t_m} - \hat{X}_{g,t_m}\|^2 / \sum_{m=2}^{M} \|X_{g,t_m}\|^2 \tag{8.4}$$

où $\hat{X}_{g,t_m}$ est l'expression prédite du gène g au temps t_m en utilisant la matrice de coefficients estimée A de chaque modèle. L'erreur été calculée pour tous les modèles, et les résultats sont présentés dans le tableau 8.3.

Modèle	Erreur NMSE
7.1	0.5094
7.5	0.0588
7.6	0.0257
7.7	0.0257
7.8	0.0247
15.1	0.0222

Table 8.1: Erreur de prédiction de 6 différent modèle pour regarder l'effet de choix de la valeur représentante de chaque classe.

On retrouve une erreur de prédiction assez mauvaise pour le modèle 7.1, ce qui est normal, car comme dans la sous-section précédente l'expression de chaque gène est ici assimilée à la moyenne de sa classe, pour une écriture non différentielle sans terme b constant pris en compte. Pour les modèles qui suivent, l'erreur de prédiction est clairement améliorée. Au final, le modèle 15.1 qui respecte les statistiques d'ordres, sur la base d'une hypothèse biologique forte, est le modèle qui minimise l'expression de l'erreur.
Cette étude de comparaison d'erreurs nous conduit donc à travailler avec ce dernier modèle (15.1) pour l'analyse biologique pour décrire l'effet de simulation sonore sur les plantes. La différence entre les différentes erreurs reste néanmoins marginale, et une étude comparative plus poussée pourrait être conduite pour détecter les différences exactes entres ces expressions.

Une représentation graphique de l'expression prédite de 3 gènes parmi les 9954 gènes obtenue en utilisant le modèle 15.1, est présentée dans la figure 8.4.

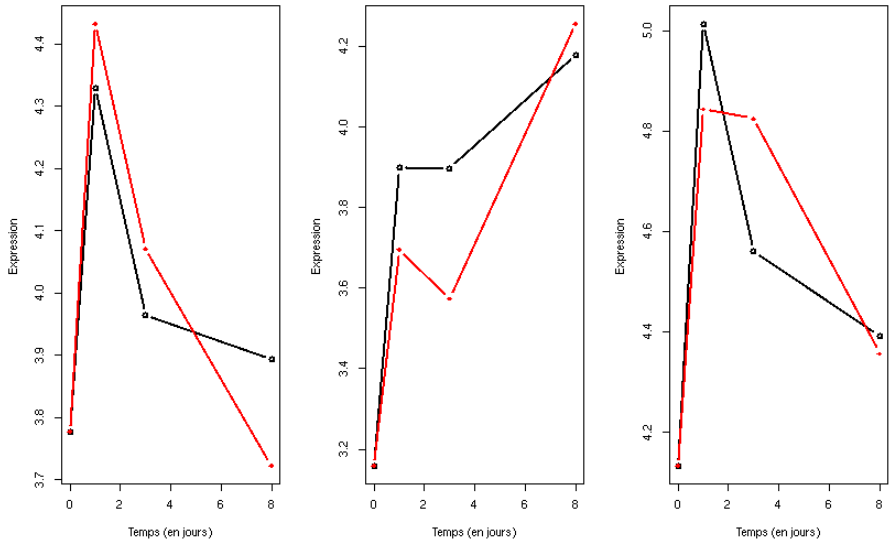


Figure 8.4: Comparaison entre l'expression réelle (noir) et l'expression prédite (rouge) pour 3 gènes parmi les 9954 gènes.

8.3 Influence des paramètres structuraux du modèle sur données simulées

L'objectif de cette section est de tester l'influence de la structure du modèle sur la qualité de l'inférence du réseau, sur la base de données simulées. Nous ferons varier le nombre des noeuds et de mesures temporelles pour des réseaux aléatoires et évaluerons la performance de l'inférence avec ou sans prise en compte explicite de l'effet direct b. Plus précisément, l'objectif de cette partie est en particulier de pouvoir mieux répondre aux deux questions suivantes :

- Quel ratio entre le nombre de classes et le nombre de mesures temporelles est le plus favorable à la bonne inférence du réseau ?
- Quelle est l'influence sur l'inférence du réseau de la prise en compte explicite d'un terme constant, qui modéliserait l'effet direct du son sur le réseau ?

8.3.1 Simulation d'expression de gènes

Dans cette section, nous utilisons un processus de simulation pour remplacer des données réelles. Le but principal de simulation est d'avoir des données aussi proches

que possible de nos données réelles (en termes des caractéristiques statistiques ; moyenne et écart-type) qui vont nous permettre de choisir le nombre de mesures temporelles et de noeuds le plus adéquat pour le réseau de réponse au son.

Nous avons choisi de simuler des trajectoires représentant les expressions de moyennes de clusters de gènes au cours du temps. Le modèle de simulation que nous avons retenu est :

$$X_{t+1} = X_t + AX_t + b + \epsilon, \tag{8.5}$$

où $A = (a_{i,j}), i, j \in \{1, ..., p\}$ est la matrice décrivant le réseau de dimension $p \times p$, b un vecteur de $\mathbb{R}^p$ décrivant un effet fixe d'une perturbation extérieur, et ϵ un processus Gaussien centré de variance σ^2 qui sera fixée à 0.01 dans tout ce chapitre.

Générer des données simulées revient donc tout d'abord dans notre cas à fixer les paramètres n et p, puis à générer aléatoirement une matrice A et un vecteur b. Une trajectoire est alors obtenue en fixant l'état initial X_0 puis en simulant aléatoirement l'équation 8.5.

Génération de la matrice A :
Nous générons une matrice aléatoirement, imposant plusieurs types de structures :

- Structuration par hubs, c'est à dire que la matrice contient quelques coefficients hautement connectés aux autres coefficients.
- Structuration creuse, c'est à dire que la matrice comporte majoritairement des coefficients non nuls.
- Structuration hypercreuse, c'est à dire que la matrice est composée principalement de coefficients nuls, voire peut être entièrement nulle.

La différence entre les trois types de matrice est basée essentiellement sur le nombre de coefficients non nuls. Un exemple est présenté dans la figure 8.5, dans cet exemple le nombre des noeuds (ou classes) est fixé à 6. Un exemple de génération de matrices aléatoires est présenté dans la figure 8.6, le nombre de noeuds pour chaque structure varie entre 20 et 60 et le nombre de liens dépend fortement du type de la matrice. Les coefficients non nuls de chaque matrice sont tirés selon une loi uniforme entre -0.5 et 0.5 (nous avons choisi le même intervalle de variation que celui observé dans l'estimation des matrices de réponse de la plante au son à partir des données réelles). À l'issue de cette étape, nous obtenons une matrice notée $A = (a_{i,j}), i, j \in \{1, ..., p\}$ avec p le nombre de classes de gènes.

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -0.052 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0.188 & 0 & -0.209 & -0.125 & 0.282 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

(a) Matrice hub

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0.067 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.400 & 0 & 0 & 0 \\ 0.453 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

(b) Matrice creuse

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -0.449 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

(c) Matrice hypercreuse

Figure 8.5: Illustration des trois types de matrices hub, creuse et hypercreuse dans le cas particulier ou le nombre de noeuds est égale à 6. La différence entre les trois matrices est liée au nombre et à la position des coefficients nuls dans chaque cas.

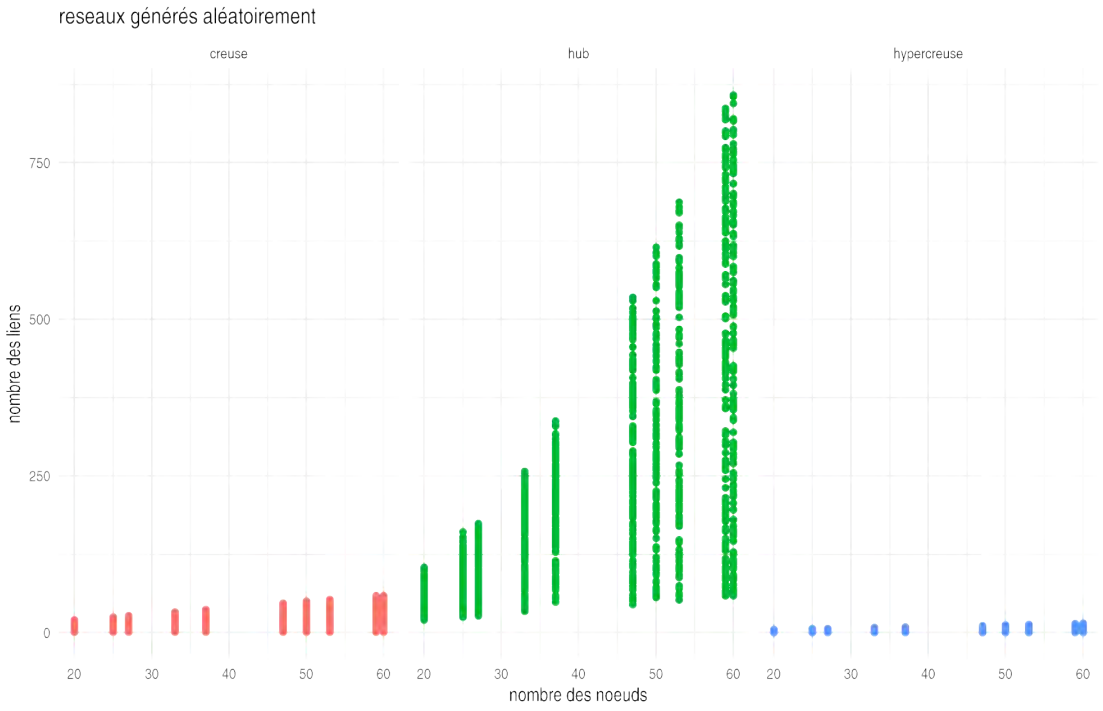


Figure 8.6: Exemple des matrices générées aléatoirement, le plus grand nombre de liens est associé aux matrices qui contiennent des hub.

Simulation du terme constant b :
Nous nous sommes basés pour générer le b sur la valeur prédite par le modèle 15.1 inféré avec 8 classes sur les données réelles. Pour adapter la taille de b au nombre de noeuds, nous avons effectué une interpolation linéaire sur le domaine $[-0.38, 0.28]$.

Simulation de X_0 :
Nous avons supposé que X_0 suit une loi normale avec une moyenne et une variance qui était fixée en utilisant les caractéristiques de notre jeu de données de son transformé en log(CPM), $X_0 \sim N(-0.02, 0.4)$.

8.3.2 Inférence et évaluation des réseaux obtenus

À partir d'une trajectoire simulée pour p clusters et n instants de mesure, nous avons utilisé un lasso classique implémenté dans le package R **glmnet** pour réaliser l'inférence des paramètres A et b vérifiant :

$$X_t = X_{t-1} + AX_{t-1} + b + \epsilon_t \tag{8.6}$$

Afin de tester l'influence de la prise en compte de b, nous avons également cherché à inférer la seule matrice A vérifiant

$$X_t = X_{t-1} + AX_{t-1} + \epsilon_t. \tag{8.7}$$

Pour évaluer les différents réseaux inférés, nous avons utilisé les scores de précision et de recall (rappel). La précision décrit la proportion de liens pertinents parmi l'ensemble des liens proposés comme pertinents par le modèle utilisé. Le recall (sensibilité) décrit la proportion des liens reconnus comme pertinents parmi l'ensemble des liens pertinents de la matrice de départ.

L'expression de précision s'exprime comme :

$$precision = \frac{VP}{VP + FP} \tag{8.8}$$

et le recall comme :

$$recall = \frac{VP}{VP + FN} \tag{8.9}$$

avec:

- VP = les vrais positifs
- FP = les faux positifs

- FN = les faux négatifs

8.3.3 Variation des nombres de noeuds et de mesures temporelles

Nous avons fait varier le nombre de noeuds, de points temporels et le type des matrices pour étudier l'effet de chaque propriété sur le réseau final.

Le nombre de noeuds (p) a été tiré aléatoirement entre 20 et 50 noeuds, ce nombre de noeuds a été fixé pour ce rapprocher du nombre de noeuds utilisés pour l'inférence du réseau sur les données réelles. Le nombre de mesures temporelles (n) a été fixé pour chaque réseau comme le nombre des noeuds ($n = p$), le nombre des noeuds divisés par deux ($n = p/2$) et le nombre des noeuds divisé par 3 ($n = p/3$). Dans les résultats suivants, pour chaque paramétrage (n, p) l'inférence a été répétée 150 fois aléatoirement.

Inférence sans prise en compte du terme constant b

Le modèle utilisé dans cette partie est le modèle sans le terme constant (équation 8.7). Les valeurs moyennes de recall sont présentées dans la figure 8.7 et celles de précisions sont présentés dans la figure 8.8.

Nous constatons que les taux de recall et précisions sont très faibles pour tous les types des matrices et pour les différents nombres de noeuds. Le taux de recall maximal pour les matrices creuses recall est de 3% et pour les hypercreuses est de 0.2%. Ce modèle estime donc très mal la matrice A avec n'importe quel nombre de noeuds et de mesures. Pour le taux de précision, ce taux était inférieur à 1% pour les matrice hubs et inférieur à 3% pour les matrices creuses. Trouver un bon compromis entre le nombre de noeuds et des mesures est alors difficile.

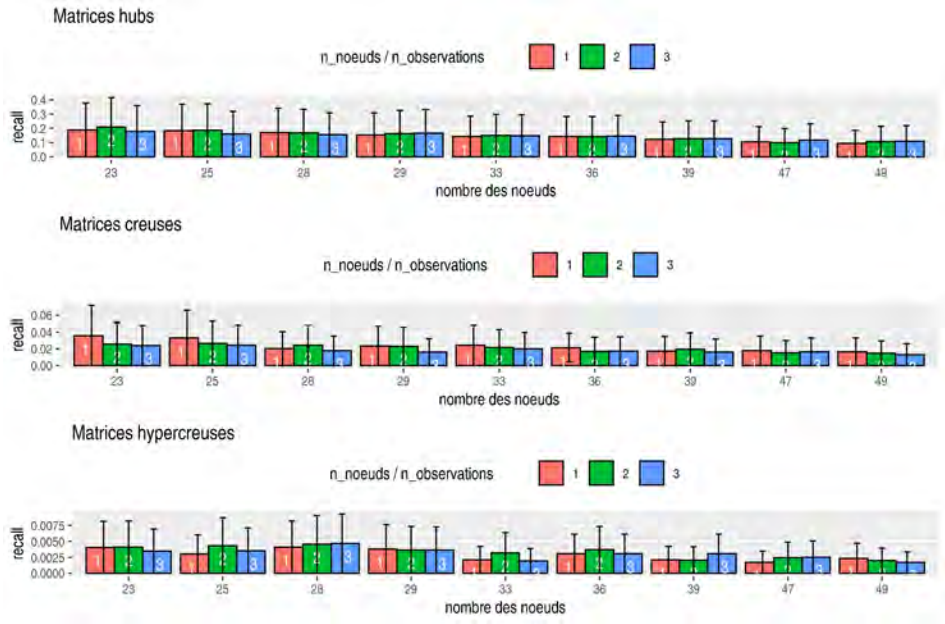


Figure 8.7: Recall, moyenne et écart-type sur 150 répétitions pour différents nombres de noeuds et de points temporels pour chaque type de matrice.

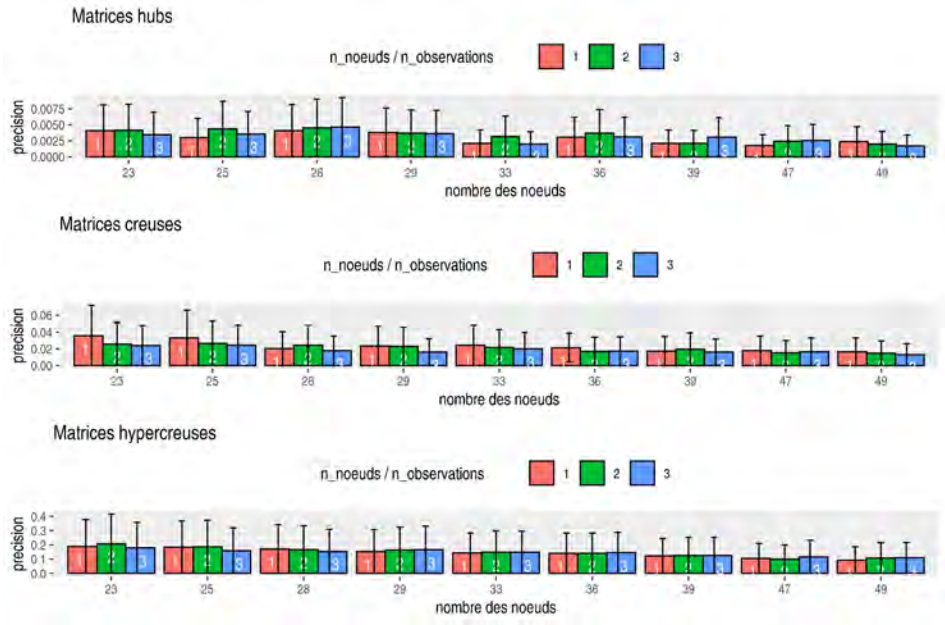


Figure 8.8: Précision, moyenne et écart-type sur 150 répétitions pour différents nombres de noeuds et de points temporels pour chaque type de matrice.

Inférence avec prise en compte du terme constant b

Le modèle utilisé est maintenant le modèle linéaire avec un terme constant qui décrit l'effet d'une perturbation extérieur sur le système de la plante (equation

8.6).

Dans cette partie, nous continuons à faire varier le nombre des noeuds entre 20 et 50 et le nombre de mesures temporelles par rapport au nombre des noeuds. Les résultats sont présentés figures8.9 et 8.10. Les taux de recall et précisions restent faibles pour tous les types des matrices et pour les différents nombre des noeuds mais en comparant ces taux avec ceux obtenus avec le modèle sans terme b on constate que l'estimation est meilleure en utilisant ce modèle. De manière évidente, les taux de recall et de précision sont améliorés avec des réseaux de petite taille (de 20 à 30), avec un nombre de mesures temporelles au moins égal au nombre des noeuds pour les matrices hubs et creuses et au nombre de noeuds sur deux pour les matrices hypercreuses.

Quel que soit le type de la matrice, si on veut estimer la matrice des coefficients d'une matrice qui contient 27 noeuds et 8 mesures temporelles, on prévoit que le taux de recall ne dépasse pas 30% (recall maximal pour les matrices creuses) et le taux de précision est assez faible et ne dépasse pas 20% (précision maximale pour les matrices hypercreuses). Enfin, on peut remarquer qu'avec n'importe quel type de matrice et pour les différents nombres de mesures temporelles, ce modèle apporte une meilleure estimation de la matrice A qu'avec l'approche précédente sans prise en compte du terme constant b.

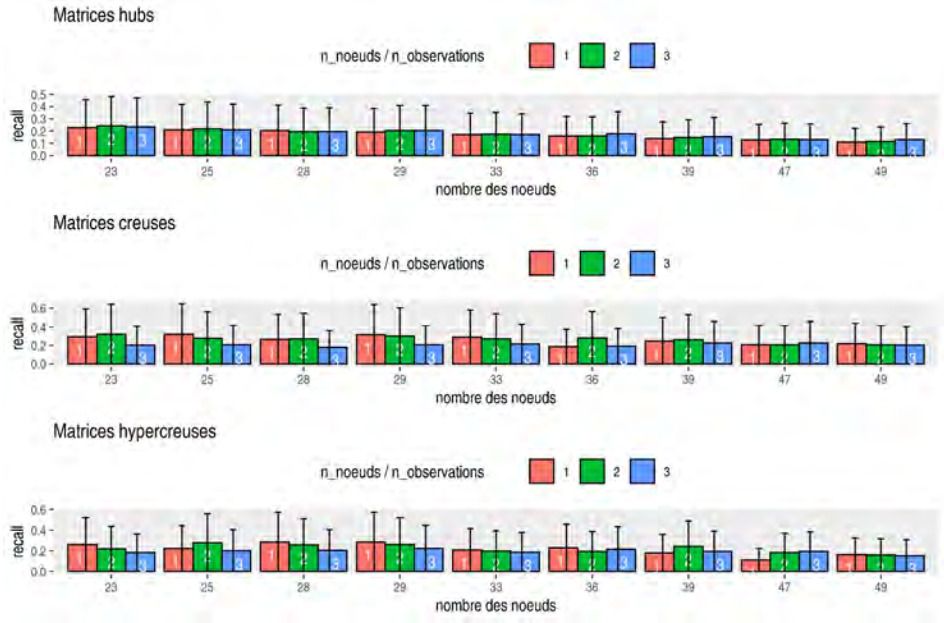


Figure 8.9: Recall, moyenne et écart-type sur 150 répétitions pour différents nombres de noeuds et de points temporels pour chaque type de matrice.

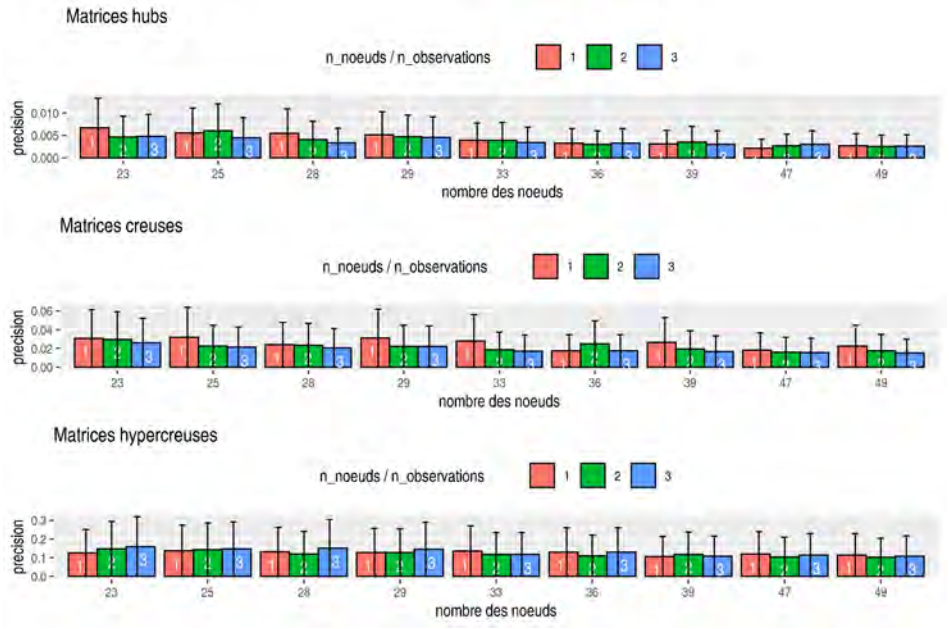


Figure 8.10: Précision, moyenne et écart-type sur 150 répétitions pour différents nombres de noeuds et de points temporels pour chaque type de matrice.

Étude autour de 8 mesures temporelles

Nous avons ici fixé le nombre de mesures temporelles à 8 et fait varier le nombre de noeuds à 8, $8 * 2$, $8 * 3$ et $8 * 4$. Le but est de nous rapprocher des paramètres d'une inférence à base de 3-signatures avec 27 noeuds et 8 mesures temporelles. Les résultats sont présentés dans la figure 8.11. Pour les matrices à hubs ou creuses, le recall peut atteindre 40% pour 8 noeuds, et s'effondre sur les matrices hypercreuses. Pour la précision, on atteint 15% pour les matrices à hubs, sans trop d'influence du nombre de noeuds, et les résultats décrochent pour les deux autres types de matrice, sauf dans le cas des matrices creuses avec 8 noeuds.

Nous pouvons également nous intéresser à la qualité de l'estimation du terme constant b. En fixant également le nombre de mesures temporelles à 8 et en faisant varier le nombre de noeuds à 8, $8 * 2$, $8 * 3$ et $8 * 4$, nous avons calculé un taux d'erreur égal à la racine de l'erreur quadratique moyenne (RMSE) de l'estimation, normalisée par la norme de b :

$$err = \frac{\|b - \hat{b}\|}{\|b\|} \tag{8.10}$$

Nous constatons (Fig. 8.12) que l'erreur varie marginalement entre les différents estimation de b et n'a pas dépassé les 18% pour les 3 types de matrices, avec maintenant la plus grande erreur relative associés au réseau contenant 32 noeuds.

En conclusion, cette évaluation par simulation met en évidence des taux de recall et de précision relativement faibles, mais toujours meilleurs lorsque la méthode d'inférence prend en compte un terme constant b. Les deux paramètres nombre de noeuds et ratio nombre de noeuds / nombre de mesures temporelles influencent négativement le recall et la précision des termes de la matrice A, avec il semble un effet plus marqué du nombre de noeuds. L'influence de ces paramètres semble moins marquée pour l'estimation du terme constant b. Dans le cas des matrices à hubs, pour 8 mesures temporelles la précision reste acceptable jusqu'à 32 noeuds.

Il semble ainsi préférable dans le cas de 8 points temporels de choisir l'inférence d'un réseau à 8 noeuds (au plus). Néanmoins, dans notre cas, les noeuds représentent des classes de gènes et travailler avec un petit nombre de classes augmente également la variance intra-classe, et donc l'erreur de prédiction. Nous en concluons alors qu'il peut être judicieux de travailler avec une classification de type 3-signatures, qui implique un maximum de 27 noeuds, afin de minimiser la variance intra-classe tout en conservant un ratio nombre de noeuds / nombre de points temporels acceptable pour le recall et la précision.

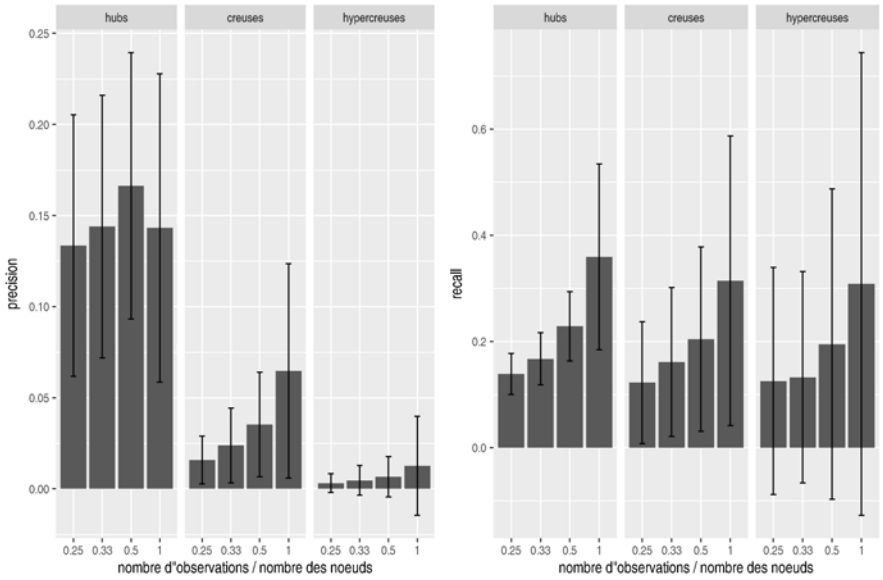


Figure 8.11: Précision et recall, moyenne et écart-type pour 8 points temporels sur 150 répétitions pour différents nombres de noeuds pour chaque type de matrice.

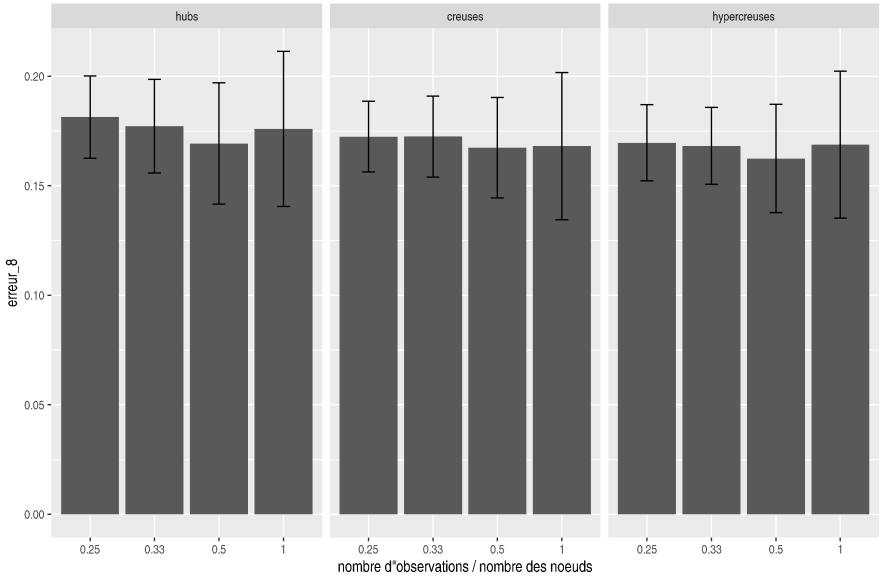


Figure 8.12: Erreur de prédiction, moyenne et écart-type pour 8 points temporels sur 150 répétitions pour différents nombres de noeuds pour chaque type de matrice.

8.4 Évaluation de la méthode de classification

Comparaison avec des méthodes de référence

Pour évaluer notre méthode de classification à base des 2−signatures, nous avons comparé les résultats obtenus par notre méthode avec les résultats de classification obtenus avec des méthodes de référence. Dans toutes les méthodes de classification évaluées, le nombre de classes K a été déterminé en utilisant le coefficient de silhouette [Rousseeuw, 1987], à l'exception de la méthode de classification à 3−signatures pour laquelle nous avons directement utilisé le nombre de classes obtenu par la méthode, soit 27 classes. Nous avons comparé les méthodes avec 4 métriques classiques utilisées pour évaluer la qualité de classification (Table 8.2) : les distances moyennes (D) entre et au sein des classes ; l'indice Davies-Bouldin (DB) [Davies and Bouldin, 1979] qui décrit la moyenne du rapport maximal entre la distance d'un point au centre de son classe et la distance entre deux centres de classe ; le coût de calcul pour chaque méthode (T) et le coefficient de silhouette (Sil) (Table 8.2). Pour cette étude de comparaison nous avons utilisé les données "son" présentées dans le chapitre 5, séries temporelles très courtes composées simplement de quatre points temporels. Pour ces données X_{g,t_m}, $g = 1, \cdots, N$ avec $N = 9954$, et $m = 1, \cdots, M$ avec $M = 4$.

Méthode	K	D_{within}	$D_{between}$	DB	Sil	T (s)	R package
2−signatures	40	0.41	2.30	0.96	0.24	0.03	
3−signatures	27	0.60	2.20	1.08	0.19	0.05	
k-means pour données fonctionnels	29	0.60	2.32	1.23	0.20	696.54	fda.usc [Febrero-Bande and de la Fuente, 2012]
Hiérarchique pour les séries temporelles	46	1.29	4.28	1.42	0.23	19.669	dtwclust [Sarda-Espinosa et al., 2018]
k-means pour données longitudinal	26	0.52	2.31	5.28	0.32	13.653	kml [Genolini et al., 2015]
k-means	28	0.51	2.31	1.09	0.32	0.044	stats [Team et al., 2018]
Hiérarchique	47	0.45	2.30	6.56	0.25	4.293	stats [Team et al., 2018]

Table 8.2: Validation des méthodes 2−signature et 3−signature avec d'autres méthodes de classification classiques. K représente le nombre de classes, D_{within} indique la distance au sein d'une classe, $D_{between}$ la distance entre les clusters, DB est l'indice de Davies-Bouldin, Sil représente la largeur moyenne de la silhouette, et T correspond au temps de calcul (en secondes).

Le nombre de classes obtenu avec la méthode de 2−signatures est similaire au nombre de classes obtenu par classification hiérarchique pour données temporelles et classification hiérarchique classique. Cette méthode a été conçue pour minimiser la distance entre les sous-classes sur la base des coefficients de silhouette. Ainsi, la distance entre les classes et le coefficient de silhouette associés à notre méthode étaient, comme attendu, parmi les plus faibles. Le nombre de classes obtenu en utilisant la méthode des 3-signatures a été déterminé sans recourir à l'indice de silhouette, C'est pourquoi le coefficient de silhouette le plus élevé est associé à cette méthode (voir tableau 8.2).

Effet de la méthode de classification sur l'inférence réseau

Pour illustrer l'effet de la classification sur l'inférence de réseau, nous avons calculé un réseau pour chaque classification produite par les méthodes précédentes, et nous avons alors comparés différents indicateurs issus de ces réseaux.

Pour l'inférence nous avons fait le choix de considérer le modèle simple autorégressif 7.1 et la méthode du package R **SIMONE** (voir section 4.2.1), sur la base des données moyennées par clusters $\bar{X}_{k,m} = \frac{1}{N_k} \sum_{g \in C_k} X_{g,t_m}$, $k = 1, \cdots, K,\ m = 1, \cdots, M$.

Nous avons ainsi pu estimer pour chaque classification (caractérisée entre autre par son nombre de classes K, et le nombre maximal de gènes au sein d'une classe $\#C_{max}$) une matrice (creuse) $A = (a_{ij})_{i,j=1,K}$, où chaque coeeficient $a_{i,j}$ represente l'influence du cluster C_j sur le cluster C_i, selon le modèle

$$\bar{X}_{m+1} = A\bar{X}_m + \epsilon_m. \tag{8.11}$$

Pour évaluer et comparer ces réseaux, des métriques de graphe couramment utilisées ont été calculés, comme le nombre de nœuds (K) et d'arêtes (N_e), le nombre de nœuds isolés (N_{iso}), les degrés de nœud maximum (d_{max}) et moyen (d_{mean}).

Nous avons également calculé l'erreur de prédiction quadratique moyenne normalisée ($NMSE$) associée à chaque classification et matrice estimée A, selon l'équation 15.4 décrite plus haut.

Méthode	K	$\#C_{max}$	N_e (% of K^2)	N_{iso} (% of K)	d_{max}	d_{mean}	err
2−signatures	40	652	36 (2%)	8 (20%)	6	2.25	0.8938
3−signatures	27	2294	26 (3%)	4 (14%)	14	1.91	0.9028
k-means pour données fonctionnels	29	908	20 (2%)	7 (24%)	7	1.81	0.9380
Hiérarchique pour les séries temporelles	46	1188	32 (1%)	14 (30%)	7	2	0.9494
k-means pour données longitudinal	26	629	20 (2%)	6 (23%)	4	2	0.9334
k-means	28	562	32 (4%)	1(3%)	5	2.37	0.9169
Hiérarchique	47	970	39 (0.1%)	13 (27%)	7	2.29	0.9358

Table 8.3: Effet de la méthode de classification de gènes sur les réseaux inférés. K est le numéro de nœud, égal au nombre de clusters, $\#C_{max}$ est la taille maximale du cluster, N_e est le nombre d'arêtes, N_{iso} est le nombre de nœuds isolés, d_{max} et d_{mean} sont les degrés maximum et moyen des nœuds, err est l'erreur quadratique moyenne.

Les réseaux inférés présentent un nombre réduit d'arêtes, de l'ordre de 2% du nombre maximal d'arêtes (K^2). Les réseaux inférés ont été trouvés principalement linéaires ($d_{mean} \approx 2$) mais contenant des hubs majeurs qui peuvent atteindre un degré maximum qui est proche de 14, sauf pour les méthodes de classification k-means et k-means pour les données longitudinales. Le réseau obtenu en utilisant la méthode de 2−signatures présente un nombre élevé de nœuds isolés, égal à 20% du nombre total de noeuds et un nombre faible de liens entre les différentes classes qui est égal à 2%. Le réseau obtenu en utilisant la méthode de 3−signatures présente 4 nœuds isolés (14% du nombre total de nœuds) et aussi un nombre faible de liens entre les différentes classes qui est égal à 3% (tableau 8.3).

L'erreur de prédiction reste relativement élevée pour tous les réseau inférés, ce qui peut s'expliquer par le fait que le modèle linéaire que l'on a ici cherché à estimer (Eq. 8.11) ne tient pas compte de l'espacement irrégulier entre les mesures, ni de la variabilité d'expression des gènes au sein des classes, et ne distingue pas l'effet du son de la dynamique naturelle de l'expression des gènes, contrairement aux modèles 7.5 à 15.3 plus élaborés.

8.5 Conclusion

Ce chapitre nous aura servi principalement à valider expérimentalement nos choix de méthodes, sur les données réelles mais aussi simulées. Pour inférer le réseau de la mémoire transcriptionnelle associées aux stimulations acoustiques répétées, on choisit alors de retenir :

- La classification à 3 signatures. Au delà du bon compromis en termes de nombre de noeuds qu'elle procure, nous retenons cette méthode de classification car elle permet de mieux gérer les paliers d'expression entre deux instants de mesure. En effet, en utilisant l'analyse transcriptomique nous avons montré que le nombre de gènes différentiellement exprimés dépend du nombre d'expositions aux stimulations sonores (chapitre 6), et donc un gène peut être ou ne pas être différentiellement exprimé entre deux instants de mesures, ce que traduit correctement l'utilisation du troisième signe "=" dans cette méthode de classification.
- L'inférence en utilisant le modèle 15.1. Ce choix modélise finement la dynamique des classes en intégrant un effet gène. Il représente explicitement l'effet direct sur cette dynamique de la stimulation sonore. Enfin, il présente l'erreur de prédiction le plus faible.

9 Inférence du réseau de la mémoire transcriptionnelle associée aux stimulations acoustiques répétées

Sommaire

9.1 Classification à 3−signatures

Dans la méthode des 3−signatures (Section 7.2.2), on associe le signe "=" au changement d'expression d'un gène entre deux instants consécutifs lorsque ce gène n'est pas différentiellement exprimé entre ces deux instants. Parmi les 9954 gènes différentiellement exprimés par rapport à la plante saine non stimulée, on en trouve 2450 qui ne sont différentiellement exprimés au cours d'aucune des 3 transitions (classe "==="), et donc 7104 répartis sur les autres classes. Pour $M = 4$ points temporels, le nombre de classe généré par la méthode est égal à 27. Pour l'inférence

de réseau, nous choisissons de de pas considérer les gènes de la classe "===", qui sont par construction des gènes peu modifiés d'un instant à l'autre, et nous aboutissons donc à une répartition de 7104 gènes parmi 26 classes, tel que décrit dans le tableau 13.1.

Ces 26 classes obtenues présentent des effectifs hétérogènes. Seize classes contiennent moins de 100 gènes, huit classes entre 101 et 1 000 gènes, et deux grandes (9 et 18) contiennent plus de 1 000 gènes.

Les signes associés à chaque groupe ainsi que le nombre des gènes sont présentés dans le tableau 13.1 :

Classe	Signe	Nombre de gènes
1	- - -	2
2	- - +	2
3	- - =	13
4	- + -	104
5	- + +	4
6	- + =	340
7	- = -	109
8	- = +	83
9	- = =	2058
10	+ - -	20
11	+ - +	91
12	+ - =	507
13	+ + -	8
14	+ + +	1
15	+ + =	9
16	+ = -	80
17	+ = +	140
18	+ = =	2294
19	= - -	6
20	= - +	70
21	= - =	310
22	= + -	98
23	= + +	1
24	= + =	88
25	= = -	379
26	= = +	287

Table 9.1: Représentation des signes de chaque classe et le nombre de gènes associés.

Aucune des 26 classes n'est spécifiquement composée de gènes associés à la MTI ou de facteurs de transcription (TF). De la même manière, les gènes associés au priming sont répartis dans 50% des classes.

Des enrichissements significatifs en processus biologiques sont présents dans 15 classes, décrivant les modifications du fonctionnement de la plante associées au traitement par des stimulations acoustiques répétées (Figure 9.1.A). Parmi ces enrichissements, on trouve la synthèse des protéines de défense, des altérations structurelles telles que l'organisation de la paroi cellulaire (classe 6) et des mouvements dépendants des microtubules (classes 4, 22) centraux pour la MTI et la défense des plantes.

9.2 Inférence du réseau descriptif de la mémoire transcriptionnelle

Pour mieux comprendre l'effet de répétitions de stimulations acoustiques, nous inférons un réseau dynamique en utilisant les 26 classes obtenues. Pour l'inférence de réseau, comme expliqué au chapitre 7, nous utilisons le modèle de dynamique 15.1, décrit par l'équation :

$$\begin{aligned} X_{g,t+1} - X_{g,t} &= A_{i,i}X_{g,t} + \sum_{j\neq i} A_{i,j}X_{j,t}^{\alpha_{g,t}} + b_i + \epsilon_t \\ X_{g,1} - X_{g,0} &= b_i + \epsilon_0, \quad \forall t = 0, \cdots 7, \quad i = 1, \cdots 26, \quad g \in C_i, \end{aligned} \tag{9.1}$$

Avec dans cette formule $\alpha_{g,t} = Pr[X_{g',t} \leq X_{g,t},\ g' \in C_j]$ l'ordre du quantile $X_{g,t}$, et $X_{j,t}^{\alpha}$ le quantile d'ordre α de la distribution des expressions des gènes $X_{j,t}$ au temps t. Les données $X_{g,t}$ entre deux observations sont reconstruites par interpolation linéaire, et l'inférence est faite classe par classe pour estimer séparément chaque ligne A_i de A et composante b_i de b, selon une méthode d'estimation LASSO (voir section 7.4.2). Pour les clusters de 19 à 26 qui contiennent des gènes qui ne sont pas significativement exprimés après la première stimulation acoustique, nous imposons dans l'inférence une valeur nulle de b (Figure. 9.1.A , Figure 9.2).

Le réseau dynamique obtenu (Figure. 9.1.B) comprend 26 noeuds inter-connectés par 184 liens orientés, révélant une organisation peu hiérarchique. Le réseau est hautement inter-connecté avec en moyenne 7 liens par noeud. Le réseau présente trois noeuds terminaux (12, 20, 24) non impliqués dans la régulation d'autres gènes et 1 noeud non régulé par d'autres (23). Les clusters associés à la perception de stimuli acoustiques sont répartis sur 18 des 26 clusters, ce qui suggère que la

perception n'est pas réalisée par des classes spécifiques (Figure. 9.1.A).

L'expression du vecteur b prédite par le modèle, décrit l'effet des stimulations acoustiques sur chaque classe. Les stimulations acoustiques répétées affectent les expressions des gènes dans chaque groupe, régulant positivement les gènes des classes de 1 à 9 et régulant négativement les gènes des classes de 10 à 18 et ne régulent pas les classes de 19 à 26.

Ensuite, nous avons déterminé les communautés existantes dans le réseau en utilisant la méthode présentée dans l'article de Blondel et al. [Blondel et al., 2008], mise en œuvre dans Gephi, un logiciel open source dédié à la visualisation et à la manipulation des réseaux [Bastian et al., 2009]. Cette méthode repose sur une approche en trois étapes bien définies. Initialement, chaque nœud du réseau est considéré comme une communauté distincte. Ensuite, la méthode évalue la qualité de la division en communautés en mesurant la modularité pour chaque nœud. Cette mesure de la modularité est utilisée comme critère d'optimisation pour déplacer itérativement les nœuds d'une communauté à une autre, cherchant à regrouper les nœuds afin de maximiser la modularité globale du réseau. Enfin, les communautés identifiées sont agrégées pour former des nœuds de niveau supérieur dans une nouvelle représentation du réseau. Ces étapes d'optimisation et d'agrégation sont répétées itérativement jusqu'à ce que la méthode ne puisse plus améliorer la modularité du réseau par de nouveaux regroupements.

Le réseau contient 4 communautés. Une communauté groupe les clusters plus liés entre eux qu'avec ceux d'autres classes. Le nombre de classes dans chaque communauté est équilibré, à l'exception de la communauté 4, qui est composée d'une seule classe (23). La communauté 2 permet la régulation de la communauté 1 par la communauté 3.

Le réseau obtenu contient trois nœuds terminaux (12, 20, 24), non impliqués dans la régulation d'autres gènes, associés à des processus tels que la réponse à la chitine et au stress endoplasmique. Cependant, la régulation entre les clusters ne converge pas vers les processus associés aux nœuds terminaux, ce qui indique que la mise en place de ces processus est complémentaire des autres processus modulés plutôt que la finalité de la mémoire transcriptionnelle.

Pour étudier l'effet dynamique du priming, nous avons quantifié la variation de l'expression des gènes entre deux stimulations acoustiques consécutives. L'expression des gènes des classes est modifiée après la première stimulation, les groupes 11 et 13 étant particulièrement sensibles, puisque l'expression des gènes de ces groupes ont varié le plus par rapport aux autres gènes d'autres groupes. L'analyse d'ontologies des classes les plus sensibles à l'effet du son (4, 11, 12, 13, 16,22) montre l'activation des processus biologiques de base tels que les processus de biosynthèse des polysaccharides, la régulation positive de la sénescence des

feuilles, la réponse aux micro-organismes, la réponse aux stimuli abiotiques, les processus basés sur les mouvements des microtubules (Figure. 9.1.C). Après le troisième stimuli, les expressions des gènes des classes 4 et 22 associés aux processus basés sur les microtubules sont fortement modulés ce qui suggère l'implication des microtubules dans la transition entre l'état naïf et primé de la plante.

Après 8 stimulations, il n'y de changements que dans l'expression des gènes de la classe 2, qui ne comprend que 2 gènes, en accord avec l'analyse phénotypique qui ne montre aucune évolution de la sensibilité des plantes à l'infection.

Ces résultats de modélisation montrent que l'activation de la mémoire transcriptionnelle associée aux répétitions de stimulations acoustiques ne se limite pas aux gènes dits de réponse, mais s'étend aux gènes présentant des profils d'expression complexes. Cette organisation non hiérarchique et complexe permet la diversification des processus de priming mobilisés chez les plantes exposées aux stimulations acoustiques.

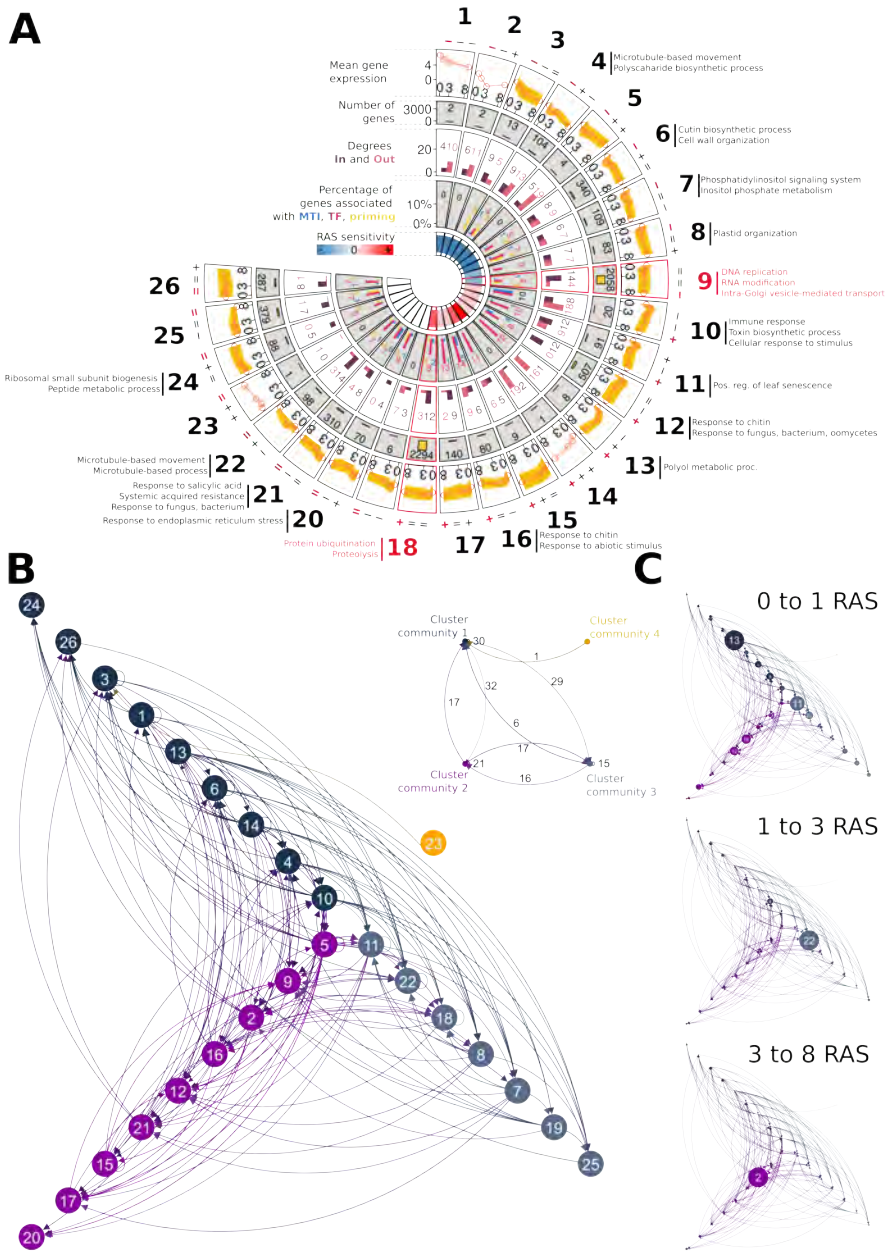


Figure 9.1: Réseau de gènes dynamique à l'échelle du génome. **A**. Propriétés des classes de gènes et leur connexion dans le réseau. **B**. Modèle dynamique d'expression des gènes organisé en 4 communautés. **C**. Variations relatives de l'expression des gènes par rapport au nombre des stimulations acoustiques.

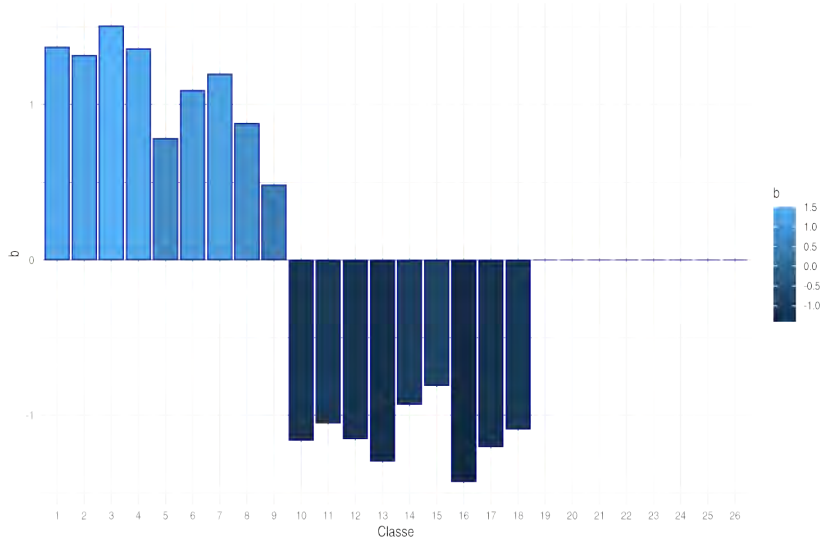


Figure 9.2: L'expression de b prédite par le modèle utilisé, qui décrit l'effet du simulation sonore sur chaque classe.

9.2.1 Robustesse du priming de la défense des plantes par des stimulations acoustiques répétées

Pour étudier plus en détail si la connectivité élevée du réseau sous-tend la robustesse de priming des défenses végétales par des stimulations acoustiques répétées, nous avons mené des expériences *in silico* simulant des délétions de gènes et évalué leur impact sur les profils d'expression des gènes.

Nous avons commencé l'analyse par la suppression séquentielle de chaque classe puis calculé l'erreur quadratique moyenne RMSE :

$$RMSE = \sqrt{\frac{1}{3}\sum_{t\in\{0,1,3\}}(X_{k,t} - \hat{X}_{k,t})^2}$$

avec $X_{k,t}$ l'expression moyenne de la classe k à l'instant t et $\hat{X}_{k,t}$ est l'expression moyenne prédite de la classe k à l'instant t, $k = 1, \ldots, 26$.

Ensuite, la RMSE associée à chaque suppression est comparée à la RMSE obtenue avec le réseau initial comprenant l'ensemble des classes. La suppression séquentielle des classes entraîne une variation limitée de la RMSE, indépendante du nombre de gènes dans chaque classe 9.3.A). Cette analyse montre que le priming de la réponse immunitaire est plutôt associée à la synergie entre les classes, plutôt qu'à une classe spécifique ou à des gènes spécifiques.

Pour valider expérimentalement la prédiction du modèle, nous avons utilisé le quintuple mutant MSL (msl4;msl5;msl6;msl9;msl10 noté $\Delta 5$) et le mutant KO MCA1. Ce choix de mutants est fait sur la base de l'analyse transcriptomique effectuée dans le chapitre 6. L'analyse précédente montrait que les gènes MSL étaient impliqués dans la mémoire transcriptionnelle et le priming des plantes par les stimulations acoustiques. AtMCA1 n'était pas modulé par les stimulations, mais était différentiellement exprimé lors de l'infection par *S. sclerotiorum*. L'analyse phénotypique de ces lignées transgéniques montre que les plantes non exposées à des stimulations acoustiques puis infectées sont plus sensibles, alors que les plantes exposées avant l'infection sont comme le sauvage, moins sensibles à l'infection (Figure. 9.3.B).

Nous avons ensuite étudié grâce au modèle l'effet du nombre de stimulations par jour sur les expressions des gènes. Cette étude montre que l'augmentation du nombre de stimulations acoustiques par jour a peu d'effet sur l'expression des gènes (Figure 9.3.C). Cependant, lorsque les stimulations sont espacées de plus d'un jour, de plus grandes variations d'expression sont observables. Pour valider les résultats du modèle, nous avons exposé les plantes d'*A. thaliana* à 2 x 3 h de stimulation acoustique par jour pendant 3 jours avant l'infection. Les plantes exposées à 2 stimuli par jour sont plus résistantes que les plantes naïves et légèrement plus sensibles que les plantes soumises à 1 stimulation journalière (Figure 9.3.D).

Ces résultats démontrent que le priming des défenses végétales par des stimulations acoustiques répétées permet de conserver le phénotype de résistance des plantes altérées dans les voies mécanoperceptives. Comme le suggère le modèle, le priming des plantes est la conséquence de l'ensemble des interactions plutôt que de gènes spécifiques. Les résultats de simulation obtenus pour des stimulations moins fréquentes posent la question de la persistance de la mémoire transcriptionnelle.

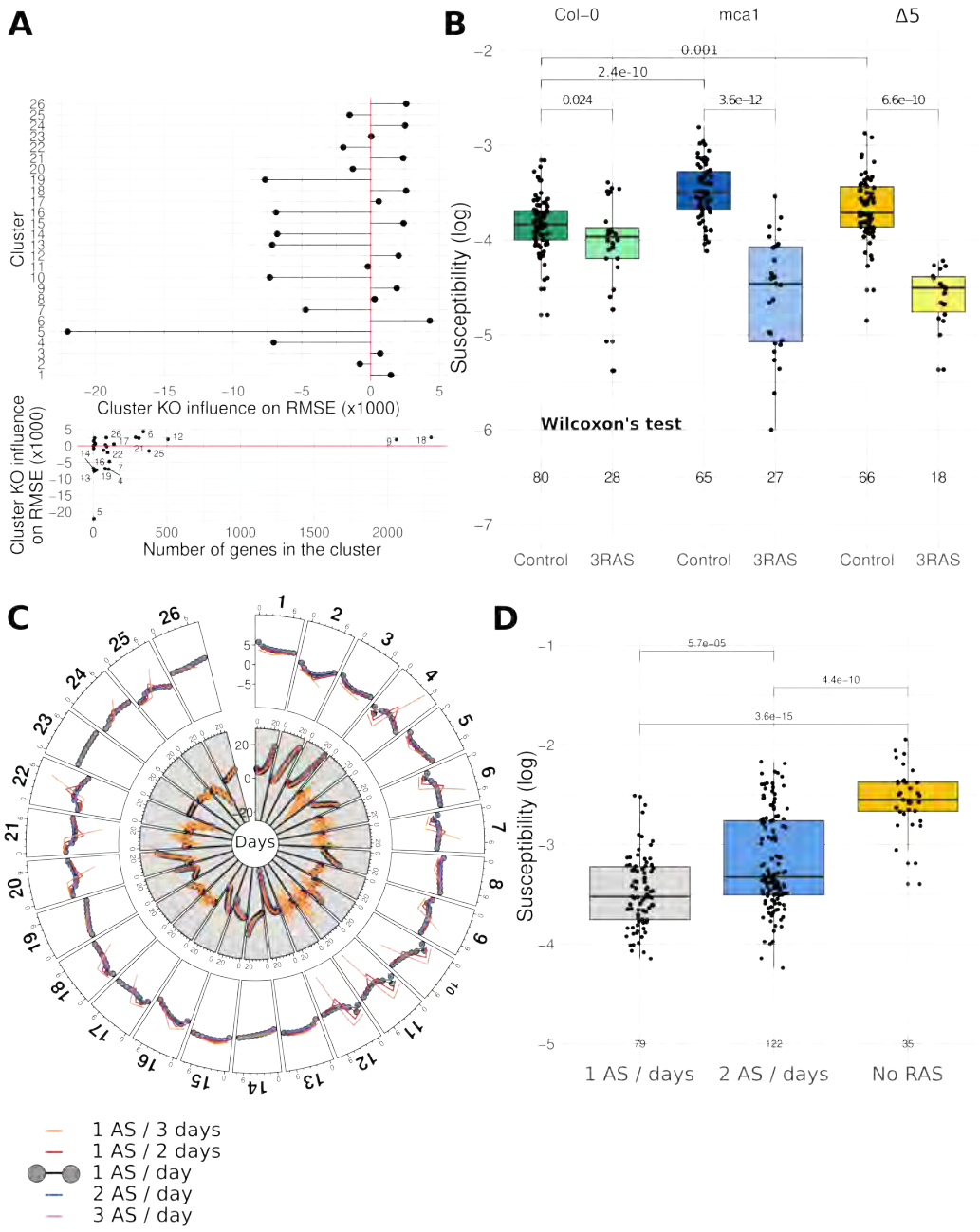


Figure 9.3: Robustesse du priming de la défense des plantes par des stimulations acoustiques. A. Simulation de la variation de l'expression génique RMSE provoquée par une délétion d'une seule classe. B. Variation phénotypique de la sensibilité des plantes contrôles et des plante exposées à des stimulations acoustiques 3 jours avant l'infection pour trois génotypes différents. C. Variations de l'expression des gènes associées à des stimulations acoustiques plus ou moins fréquentes. Variation sur 8 jours (couche externe) et 24 jours (couche interne). D. Variations phénotypiques de sensibilité à *S. sclerotiorum* pour les plantes exposées à différentes fréquences de stimulations.

9.2.2 Mémoire transcriptionnelle et stabilité temporelle du réseau

Pour caractériser la durée de la mémoire transcriptionnelle, nous étudions les valeurs propres de la matrice des coefficients A afin d'identifier les modes du système (Première méthode de Liapounov [Liapounoff, 1907, Richard and Ksouri, 2008]) et de décrire comment l'expression génique des classes évolue en l'absence de la stimulation sonore.

Chaque mode représente une combinaison linéaire des classes. L'ensemble des valeurs propres est composé de 11 valeurs propres réelles, 12 complexes et 3 nulles. 18 des 25 modes sont des modes stables avec des valeurs propres à parties réelles négatives (stabilité au sens de Liapounov). Les modes instables ont une contribution limitée aux variations des expressions géniques (Figure 9.4.A).

Le premier mode, qui contribue à 30 % de la variation de l'expression des gènes (Le module de ce mode, représente 30 % de la somme totale des modules des modes), est un mode stable et consiste en la variation de 25 classes sur 26.

Le temps caractéristique requis pour qu'un mode stable revienne à son état initial est calculé comme l'inverse de la partie réelle de sa valeur propre (Figure 9.4.B). En d'autres termes, pour un mode associé à la valeur propre $\lambda = a + ib$, le temps caractéristique pour son retour à l'état initial est $-\frac{1}{a}$. Par conséquent, plus un mode contribue à l'énergie totale, plus rapidement il revient à son état initial.

En supposant que le premier mode domine la mémoire transcriptionnelle, l'analyse des valeurs propres montre qu'une plante primée exposée à 3 stimulations "oublierait" le priming après 1.5 jours sans stimulations. Pour tester cette proposition déduite du modèle, les plantes Col-0 ont été soumises à 3 stimulations acoustiques. Ensuite, les plantes ont été infectées 0, 1, 2 et 3 jours après la dernière stimulation et on a mesuré leur sensibilité au champignon *S. sclerotiorum* (Figure 9.4.C). L'analyse phénotypique montre que l'effet du priming est visible 1 jour après la dernière stimulation, mais disparaît dès le 2éme jour sans stimulation. Ces résultats confirment l'efficacité du modèle pour prédire la durée de la mémoire transcriptionnelle de l'effet de répétitions des stimulations acoustiques (Figure 9.4.C). Cette mémoire transcriptomique persiste moins de 2 jours (1.5 jours).

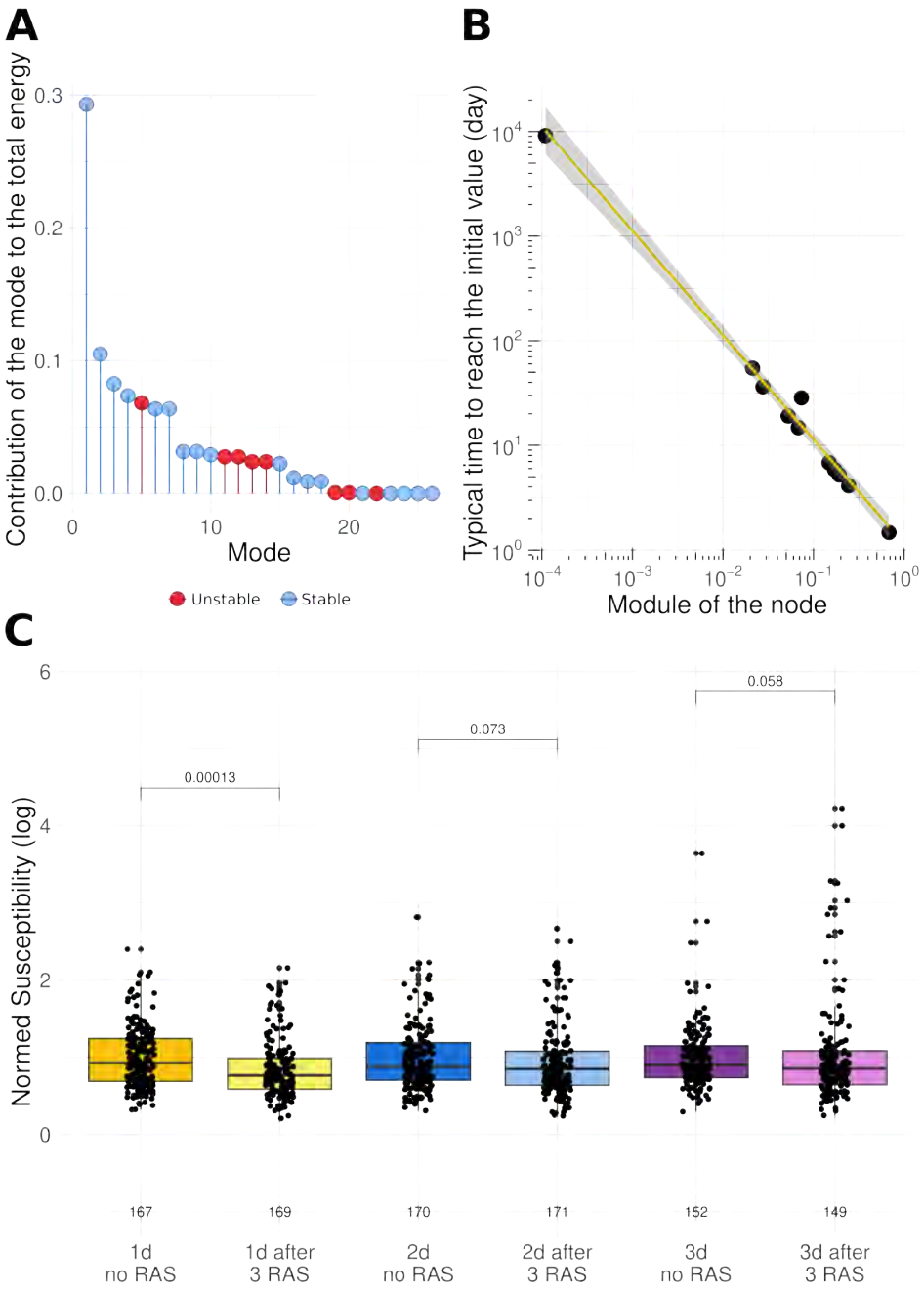


Figure 9.4: Stabilité temporelle de réseau. A. Contribution de chaque nœud à l'énergie totale. Le mode stable est en bleu, le mode instable en rouge. B. Relation entre le temps typique nécessaire à un mode pour retrouver sa valeur initiale et son énergie. C. Variations phénotypiques des plantes d'*A. thaliana* exposées à 3 stimulations acoustiques répétées et infectées 1, 2 et 3 jours après la dernière stimulations acoustiques.

9.3 Discussion

Pour comprendre l'effet des stimulations acoustiques répétées nous avons utilisé un modèle dynamique de régulation sans ajouter d'hypothèses biologiques *a priori*.

Le modèle permet la prédiction des principales caractéristiques phénotypiques émergentes de la mémoire transcriptionnelle. Cependant, ce modèle présente des limites. Les relations identifiées entre les groupes ont seulement une signification mathématique, limitant la compréhension globale des fondements biologiques régissant les interactions génétiques. Cette limitation a probablement masqué les gènes initiateurs de la mémoire transcriptionnelle, notamment ceux impliqués dans les processus de transduction. Une solution serait en conservant une approche mathématique similaire d'exploiter des données de séquençage d'ARN avec un pas de temps plus court, caractéristique des interactions génétiques. Cette approche pourrait potentiellement dévoiler les premiers mécanismes sous-jacents à la perception des ondes acoustiques par les plantes.
Les recherches sur les effets précoces des ondes acoustiques sur les plantes suggèrent qu'un pas de temps plus approprié pour de telles investigations est d'une minute, ce qui représente une résolution temporelle environ 1 000 fois plus fine que les données utilisées dans cette thèse. Appliquer le même modèle à cette résolution temporelle pourrait révéler des subtilités inaperçues jusqu'à ce jour et fournir un aperçu plus détaillé des réponses rapides et des mécanismes de régulation déclenchés par les stimulations acoustiques répétées dans les systèmes végétaux.

Un nombre croissant de chercheurs proposent qu'il n'est pas invraisemblable que la composante acoustique de la niche écologique informe la plante de son environnement ([Appel and Cocroft, 2023a, Khait et al., 2023]). Appel and Cocroft (2014) ont démontré que les plantes d'*Arabidopsis thaliana* exposées aux vibrations provoquées par la mastication de la chenille *Pieris rapae* présentaient des taux de glucosinolates plus élevés que les plantes soumises aux vibrations des *cicadelles* ou du vent [Appel and Cocroft, 2014]. Les découvertes d'Apple et de Cocroft suggèrent que la réponse de la plante pourrait être spécifique du spectre sonore. Dans notre étude, le spectre est limité à un pic à 1000 Hz. Cette fréquence est plus importante dans le spectre des *cicadelles* que chez *Pieris rapae* [Appel and Cocroft, 2014], mais elle induit une défense efficace. La compréhension de la manière dont les plantes perçoivent les ondes acoustiques reste un obstacle à une interprétation satisfaisante de ces résultats expérimentaux.

10 Conclusion générale et perspectives

Le fonctionnement des organismes vivants émerge d'une multitude d'interactions moléculaires, résultant en partie de l'expression génique. Pour mieux appréhender cette dynamique, la modélisation de la régulation de l'expression génique offre un potentiel considérable, permettant de synthétiser la complexité du vivant et d'extraire les paramètres essentiels. Les techniques actuelles de séquençage de l'ARN ouvrent la voie à la création de vastes ensembles de données décrivant l'évolution dynamique des paysages transcriptomiques, nécessitant des développements mathématiques pour concevoir des méthodes adaptées et performantes.

Au cours de cette thèse, je me suis particulièrement intéressé à l'inférence de réseaux de régulation dynamique à l'échelle du génome à l'aide des équations différentielles ordinaires, ainsi qu'à sa mise en pratique pour explorer la mémoire transcriptionnelle associée à des stimulations sonores répétées. Le très faible nombre de réplicats biologiques (3 ou 4 réplicats par condition), devant le grand nombre de gènes ($p \sim 28000$ pour *A. thaliana*), implique une première étape de diminution des dimensions, conduisant à la mise en place d'une méthode simple de classification basée sur les signes de variations des dynamiques temporelles d'expression des gènes. La comparaison de cette méthode aux méthodes classiques a permis de montrer qu'elle était la plus adaptée aux séries temporelles très courtes décrites avec un pas de temps irrégulier (chapitre 8).

Nous nous sommes intéressés ensuite à l'inférence de réseau à partir des classes obtenues. Nous avons proposé un modèle linéaire permettant de décrire l'expression de la variation de l'expression des gènes d'une classe comme la résultante proportionnelle des niveaux d'expression des gènes des autres classes plus un effet fixe décrivant une sollicitation externe du système. Nous avons montré que le critère

d'agrégation le plus adapté pour l'inférence tient compte de l'expression de tous les gènes de chaque classe ainsi que le respect des statistiques d'ordre entre les différents. Cette approche tient compte de l'expression des gènes de chaque classe afin d'intégrer la variance intra-classe (chapitre 7).

Cette méthode d'ajustement est associée à une hypothèse biologique assez forte : les gènes fortement exprimés dans une classe régulent les gènes qui ont des expressions fortes dans les autres classes. Cette hypothèse est discutable mais souvent faite implicitement en biologie moléculaire pour identifier des gènes majeurs. Par ailleurs, notre modèle ne modélise pas les interactions entre gènes d'une même classe, ce qui limite fortement la recherche des gènes clés dans le processus modélisé en vue d'une validation fonctionnelle. Une solution serait d'exploiter des données de séquençage d'ARN avec un pas de temps plus court, caractéristique des interactions génétiques.

Nous avons ensuite mené une étude comparative entre deux modèles dynamiques en utilisant des données dynamiques simulées (chapitre 8). Ce travail par simulation met en évidence des taux de recall et de précision relativement faibles, mais toujours meilleurs lorsque la méthode d'inférence prend en compte un terme constant.

Au cours de cette thèse, je me suis également intéressé à modéliser les corrélations entre l'expression des gènes et le phénotype mesuré (Annexe 4). D'un point de vue biologique, ces relations nous auraient aidé à cibler des gènes plus importants pour le processus de priming. Le modèle linéaire et la corrélation simple que j'ai testés se sont avérés infructueux en raison des fortes corrélations dans nos données et du faible nombre de mesures temporelles. Pour dépasser cette limite, l'enrichissement des données d'expression est nécessaire. On pourrait, par exemple, penser à mesurer l'expression d'un gène représentatif par classe en q-RT PCR avec un pas de temps plus fin que celui de la journée.

Nous avons volontairement conduit ce travail de modélisation sans *a priori*. Notre démarche peut néanmoins s'adapter au cas où l'on souhaiterait modéliser les régulations dynamiques d'un ensemble de gènes plus restreint choisi pour des critères biologiques.

Malgré les limites de notre approche, l'application de la méthode d'inférence au cas du priming de la réponse quantitative d'*Arabidopsis thaliana* par des stimulations acoustiques répétées nous a permis de mieux comprendre la biologie sous-jacente. Nous avons mis en évidence l'existence d'une mémoire transcriptionnelle qui mobilise plus de 9000 gènes et qui permet la mise en place d'une résistance immunitaire chez les plantes saines exposées aux stimulations, le recrutement de milliers de gènes chez la plante primée en contexte d'infection et le priming de gènes de défense (chapitre 6). Même si ce travail n'a pas permis d'établir si la mécanoperception était mobilisée pour la perception des ondes sonores, nous avons néanmoins montré

une connexion forte entre la MTI (Mechano-signalling Triggered Immunity) et le priming par les ondes acoustiques. Des études complémentaires permettraient sans doute de déterminer si la répétition des ondes sonores permet de potentialiser la MTI. Nous avons également montré qu'il était possible, à partir du réseau obtenu, de prédire les propriétés conférées à la plante par la mémoire transcriptionnelle. Nous avons aussi montré que le priming n'était pas dépendant d'un nombre limité de gènes et qu'ainsi, même chez des plantes exhibant des mutations aux effets néfastes pour l'immunité, il était possible de primer leur réponse immunitaire. De la même manière, nous avons montré que l'intensité du gain de résistance était dépendante du temps caractéristique des régulations génétiques plutôt que de celui de la cadence des stimulations sonores. Augmenter la fréquence des stimulations ne permet pas d'obtenir des plantes plus résistantes. L'analyse des modes d'expression a permis d'établir que la mémoire transcriptionnelle est relativement courte. Après 1.5 jour, il n'existe plus d'effets visibles du priming (chapitre 9). À la vue de ces résultats, l'importance des régulations épigénétiques dans la mémoire transcriptionnelle apparaît secondaire.

Références bibliographiques

A Alexa and J Rahnenfuhrer. topgo: Enrichment analysis for gene ontology. r package version 2.28. 0. *Cranio*, 2016.

Adrian Alexa, Jörg Rahnenführer, and Thomas Lengauer. Improved scoring of functional groups from gene expression data by decorrelating go graph structure. *Bioinformatics*, 22(13):1600–1607, 2006.

Genevera I Allen and Zhandong Liu. A log-linear graphical model for inferring genetic networks from high-throughput sequencing data. In *2012 IEEE International Conference on Bioinformatics and Biomedicine*, pages 1–6. IEEE, 2012.

Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Nature Precedings*, pages 1–1, 2010.

Olivia Angelin-Bonnet, Patrick J Biggs, Samantha Baldwin, Susan Thomson, and Matthieu Vignes. sismonr: simulation of in silico multi-omic networks with adjustable ploidy and post-transcriptional regulation in r. *Bioinformatics*, 36(9): 2938–2940, 2020.

Heidi Appel and Reginald Cocroft. Plant ecoacoustics: a sensory ecology approach. *Trends in Ecology & Evolution*, 2023a.

Heidi Appel and Reginald Cocroft. Ecology & Evolution Plant ecoacoustics : a sensory ecology approach. *Trends in Ecology & Evolution*, pages 1–8, 2023b. ISSN 0169-5347. doi: 10.1016/j.tree.2023.02.001. URL https://doi.org/10.1016/j.tree.2023.02.001. Publisher: Elsevier Ltd.

Heidi M Appel and Reginald B Cocroft. Plants respond to leaf vibrations caused by insect herbivore chewing. *Oecologia*, 175(4):1257–1266, 2014.

Paul L Auer and Rebecca W Doerge. A two-stage poisson model for testing rna-seq data. *Statistical applications in genetics and molecular biology*, 10(1), 2011.

Zoya Avramova. Transcriptional 'memory'of a stress: transient chromatin and memory (epigenetic) marks at stress-response genes. *The Plant Journal*, 83(1): 149–159, 2015.

Zoya Avramova. Defence-related priming and responses to recurring drought: Two manifestations of plant transcriptional memory mediated by the ABA and JA signalling pathways. *Plant, Cell & Environment*, 42(3):983–997, 2019. ISSN 1365-3040. doi: 10.1111/pce.13458. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/pce.13458. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/pce.13458.

Laura Bacete, Hugo Melida, Eva Miedes, and Antonio Molina. Plant cell wall-mediated immunity: cell wall changes trigger disease resistance responses. *The Plant Journal*, 93(4):614–636, 2018.

Thomas Badet, Ophélie Léger, Marielle Barascud, Derry Voisin, Pierre Sadon, Remy Vincent, Aurélie Le Ru, Claudine Balagué, Dominique Roby, and Sylvain Raffaele. Expression polymorphism at the arpc 4 locus links the actin cytoskeleton with quantitative disease resistance to sclerotinia sclerotiorum in arabidopsis thaliana. *New Phytologist*, 222:480–496, 4 2019. ISSN 0028-646X. doi: 10.1111/nph.15580. URL https://onlinelibrary.wiley.com/doi/10.1111/nph.15580.

Anthony Bagnall, Jason Lines, Aaron Bostrom, James Large, and Eamonn Keogh. The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data mining and knowledge discovery*, 31(3):606–660, 2017.

Michael Banf and Seung Y Rhee. Computational inference of gene regulatory networks: approaches, limitations and opportunities. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1860(1):41–52, 2017.

Mukesh Bansal, Giusy Della Gatta, and Diego Di Bernardo. Inference of gene regulatory networks and compound mode of action from time course gene expression profiles. *Bioinformatics*, 22(7):815–822, 2006.

Adelin Barbacci, Thi??ry Constant, and G??rard Nepveu. Theoretical and experimental study of a mechanical model describing the trunk behaviour of mature beech trees (Fagus sylvatica L.) under the static loading of the crown. *Trees - Structure and Function*, 23(6):1137–1147, 2009.

Adelin Barbacci, Marc Lahaye, and Vincent Magnenet. Another Brick in the Cell Wall: Biosynthesis Dependent Growth Model. *PLoS ONE*, 8(9), 2013.

Adelin Barbacci, Vincent Magnenet, and Marc Lahaye. Thermodynamical journey in plant biology. *Frontiers in plant science*, 6(June):481, 2015. ISSN 1664-462X. doi: 10.3389/fpls.2015.00481. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4485339&tool=pmcentrez&rendertype=abstract. ISBN: 1664-462X.

Adelin Barbacci, Olivier Navaud, Malick Mbengue, Marielle Barascud, Laurence Godiard, Mehdi Khafif, Aline Lacaze, and Sylvain Raffaele. Rapid identification of an arabidopsis nlr gene as a candidate conferring susceptibility to sclerotinia sclerotiorum using time-resolved automated phenotyping. *The Plant Journal*, 103 (2):903–917, 2020.

B Basavanagoud, Veena R Desai, and Shreekant Patil. () connectivity index of graphs. *Applied Mathematics and Nonlinear Sciences*, 2(1):21–30, 2017.

Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. Gephi: an open source software for exploring and manipulating networks. In *Proceedings of the international AAAI conference on web and social media*, volume 3, pages 361–362, 2009.

Renaud Bastien, Tomas Bohr, Bruno Moulia, and Stéphane Douady. Unifying model of shoot gravitropism reveals proprioception as a central feature of posture control in plants. *Proceedings of the National Academy of Sciences of the United States of America*, 110(2):755–60, 2013. ISSN 1091-6490. doi: 10.1073/pnas.1214301109. URL http://www.pnas.org/content/110/2/755.full. arXiv: 84872174159 ISBN: 1091-6490 (Electronic)\r0027-8424 (Linking).

L Benikhlef, F L'Haridon, E Abou-Mansour, M Serrano, M Binda, A Costa, and S Lehmann. i métraux, j.(2013). perception of soft mechanical stress in arabidopsis leaves activates disease resistance. *BMC Plant Biology*, 13(1):133.

Lehcen Benikhlef, Floriane L'Haridon, Eliane Abou-Mansour, Mario Serrano, Matteo Binda, Alex Costa, Silke Lehmann, and Jean-Pierre Métraux. Perception of soft mechanical stress in arabidopsis leaves activates disease resistance. *BMC plant biology*, 13(1):1–12, 2013.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

Julian Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):192–225, 1974.

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

Wang Bochu, Chen Xin, Wang Zhen, Fu Qizhong, Zhou Hao, and Ran Liang. Biological effect of sound field stimulation on paddy rice seeds. *Colloids and Surfaces B: Biointerfaces*, 32(1):29–34, 2003.

Wang Bochu, Shao Jiping, Li Biao, Lian Jie, and Duan Chuanren. Soundwave stimulation triggers the content change of the endogenous hormone of the chrysanthemum mature callus. *Colloids and surfaces B: Biointerfaces*, 37(3-4):107–112, 2004.

Richard Bonneau, David J Reiss, Paul Shannon, Marc Facciotti, Leroy Hood, Nitin S Baliga, and Vesteinn Thorsson. The inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. *Genome biology*, 7:1–16, 2006.

Janet Braam. In touch: plant responses to mechanical stimuli. *New Phytologist*, 165(2):373–389, 2005.

Hortense Brun, Anne-Marie Chèvre, Bruce DL Fitt, Stephen Powers, Anne-Laure Besnard, Magali Ermel, Virginie Huteau, Bruno Marquer, Frédérique Eber, Michel Renard, et al. Quantitative resistance increases the durability of qualitative resistance to leptosphaeria maculans in brassica napus. *New Phytologist*, 185(1): 285–299, 2010.

Atul J Butte and Isaac S Kohane. Unsupervised knowledge discovery in medical databases using relevance networks. In *Proceedings of the AMIA Symposium*, page 711. American Medical Informatics Association, 1999.

Atul J Butte, Pablo Tamayo, Donna Slonim, Todd R Golub, and Isaac S Kohane. Discovering functional relationships between rna expression and chemotherapeutic susceptibility using relevance networks. *Proceedings of the National Academy of Sciences*, 97(22):12182–12186, 2000.

Camille Charbonnier, Julien Chiquet, and Christophe Ambroise. Weighted-lasso for structured network inference from time course data. *Statistical applications in genetics and molecular biology*, 9(1), 2010.

E. Wassim Chehab, Elizabeth Eich, and Janet Braam. Thigmomorphogenesis: A complex plant response to mechano-stimulation. *Journal of Experimental Botany*, 60(1):43–56, 2009. ISSN 00220957. doi: 10.1093/jxb/ern315. ISBN: 1460-2431 (Electronic)\n0022-0957 (Linking).

E Wassim Chehab, Chen Yao, Zachary Henderson, Se Kim, and Janet Braam. Arabidopsis touch-induced morphogenesis is jasmonate mediated and protects against pests. *Current Biology*, 22(8):701–706, 2012.

Xiaohui Chen, Ming Chen, and Kaida Ning. Bnarray: an r package for constructing gene regulatory networks from microarray data by using bayesian network. *Bioinformatics*, 22(23):2952–2954, 2006.

Julien Chiquet, Alexander Smith, Gilles Grasseau, Catherine Matias, and Christophe Ambroise. Simone: Statistical inference for modular networks. *Bioinformatics*, 25(3):417–418, 2009.

Julien Chiquet, Guillem Rigaill, and Martina Sundqvist. A multiattribute gaussian graphical model for inferring multiscale regulatory networks: an application in breast cancer. In *Gene Regulatory Networks*, pages 143–160. Springer, 2019.

Bosung Choi, Ritesh Ghosh, Mayank Anand Gururani, Gnanendra Shanmugam, Junhyun Jeon, Jonggeun Kim, Soo-Chul Park, Mi-Jeong Jeong, Kyung-Hwan Han, Dong-Won Bae, et al. Positive regulatory role of sound vibration treatment in arabidopsis thaliana against botrytis cinerea infection. *Scientific Reports*, 7(1): 1–14, 2017a.

Yoonha Choi, Marc Coram, Jie Peng, and Hua Tang. A poisson log-normal model for constructing gene covariation network using rna-seq data. *Journal of Computational Biology*, 24(7):721–731, 2017b.

Donald F Cipollini Jr. Wind-induced mechanical stimulation increases pest resistance in common bean. *Oecologia*, pages 84–90, 1997.

Nicole Cloonan and Sean M Grimmond. Transcriptome content and dynamics at single-nucleotide resolution. *Genome biology*, 9(9):1–4, 2008.

Uwe Conrath, Gerold JM Beckers, Caspar JG Langenbach, and Michal R Jaskiewicz. Priming for enhanced defense. *Annual review of phytopathology*, 53:97–119, 2015.

Juliana Costa-Silva, Douglas Domingues, and Fabricio Martins Lopes. Rna-seq differential expression analysis: An extended review and a software tool. *PloS one*, 12(12):e0190152, 2017.

C Coutand and B Moulia. Biomechanical study of the effect of a controlled bending on tomato stem elongation: local strain sensing and spatial integration of the signal. *Journal of experimental botany*, 51(352):1825–42, 2000. ISSN 0022-0957. doi: 10.1093/jexbot/51.352.1825. URL http://www.ncbi.nlm.nih.gov/pubmed/11113161. ISBN: 0022-0957 (Print)\r0022-0957 (Linking).

Francis Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–563, 1970.

Peter A Crisp, Diep R Ganguly, Aaron B Smith, Kevin D Murray, Gonzalo M Estavillo, Iain Searle, Ethan Ford, Ozren Bogdanović, Ryan Lister, Justin O Borevitz, et al. Rapid recovery gene downregulation during excess-light stress and recovery in arabidopsis. *The Plant Cell*, 29(8):1836–1863, 2017.

Haitao Cui, Kenichi Tsuda, and Jane E Parker. Effector-triggered immunity: from pathogen perception to robust defense. *Annual review of plant biology*, 66:487–511, 2015.

David L Davies and Donald W Bouldin. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2):224–227, 1979.

Pierre Marc Delaux and Sebastian Schornack. Plant evolution driven by interactions with symbiotic and pathogenic microbes. *Science*, 371, 2 2021. ISSN 10959203. doi: 10.1126/SCIENCE.ABA6605/ASSET/132D3A15-53D4-4E6B-939E-F8B62CB898BA/ASSETS/GRAPHIC/371_ABA6605_F3.JPEG. URL https://www.science.org/doi/10.1126/science.aba6605.

Régine Delourme, L Bousset, Magali Ermel, Philippe Duffe, Anne-Laure Besnard, Bruno Marquer, I Fudal, Juliette Linglin, Joel Chadoeuf, and Hortense Brun. Quantitative resistance affects the speed of frequency increase but not the diversity of the virulence alleles overcoming a major resistance gene to leptosphaeria maculans in oilseed rape. *Infection, genetics and evolution*, 27:490–499, 2014.

Florent Delplace. *La résistance quantitative à Xanthomonas campestris pv campestris: la kinase atypique RKS1 contrôle un réseau immunitaire décentralisé et régule le microbiote foliaire*. PhD thesis, Toulouse 3, 2021.

Florent Delplace, Carine Huard-Chauveau, Ullrich Dubiella, Mehdi Khafif, Eva Alvarez, Gautier Langin, Fabrice Roux, Rémi Peyraud, and Dominique Roby. Robustness of plant quantitative disease resistance is provided by a decentralized immune network. *Proceedings of the National Academy of Sciences of the United States of America*, 117:18099–18109, 7 2020. ISSN 10916490. doi: 10.1073/pnas.2000078117. URL https://www.pnas.org/lookup/suppl/.

Arthur P Dempster. Covariance selection. *Biometrics*, pages 157–175, 1972.

Henri Desaint, Nathalie Aoun, Laurent Deslandes, Fabienne Vailleau, Fabrice Roux, and Richard Berthomé. Fight hard or die trying: when plants face pathogens under heat stress. *New Phytologist*, 229(2):712–734, 2021.

Atul Deshpande, Li-Fang Chu, Ron Stewart, and Anthony Gitter. Network inference with granger causality ensembles on single-cell transcriptomics. *Cell reports*, 38(6): 110333, 2022.

T Gregory Dewey. From microarrays to networks: mining expression time series. *Drug discovery today*, 7(20):s170–s175, 2002.

T Gregory Dewey and David J Galas. Dynamic models of gene expression and classification. *Functional & Integrative Genomics*, 1:269–278, 2001.

Marie Didelon. *Étude de l'impact des conditions climatiques sur la résistance quantitative de A. thaliana face au champignon pathogène S. sclerotiorum*. PhD thesis, Université de Toulouse, 2022.

Marie-Agnès Dillies, Andrea Rau, Julie Aubert, Christelle Hennequet-Antier, Marine Jeanmougin, Nicolas Servant, Céline Keime, Guillemette Marot, David Castel, Jordi Estelle, et al. A comprehensive evaluation of normalization methods for illumina high-throughput rna sequencing data analysis. *Briefings in bioinformatics*, 14(6):671–683, 2013.

Peter N Dodds and John P Rathjen. Plant immunity: towards an integrated view of plant–pathogen interactions. *Nature Reviews Genetics*, 11(8):539–548, 2010.

Timo Engelsdorf, Nora Gigli-Bisceglia, Manikandan Veerabagu, Joseph F McKenna, Lauri Vaahtera, Frauke Augstein, Dieuwertje Van der Does, Cyril Zipfel, and Thorsten Hamann. The plant cell wall integrity maintenance and immune signaling systems cooperate to control stress responses in arabidopsis thaliana. *Science signaling*, 11(536):eaao3070, 2018a.

Timo Engelsdorf, Nora Gigli-Bisceglia, Manikandan Veerabagu, Joseph F. McKenna, Lauri Vaahtera, Frauke Augstein, Dieuwertje Van der Does, Cyril Zipfel, and Thorsten Hamann. The plant cell wall integrity maintenance and immune signaling systems cooperate to control stress responses in Arabidopsis thaliana. *Science Signaling*, 11(536), 2018b. ISSN 19379145. doi: 10.1126/scisignal.aao3070.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456): 1348–1360, 2001.

Manuel Febrero-Bande and Manuel Oviedo de la Fuente. Statistical computing in functional data analysis: The r package fda. usc. *Journal of statistical Software*, 51:1–28, 2012.

Frédéric Garcia, Ophélie Léger, Rémy Vincent, Aroune Duclos, Noé Jimenez, Eric Badel, Nathalie Leblanc-Fournier, Sylvain Raffaele, and Adelin Barbacci. Thigmoimmunity: mechanical signals prime the quantitative disease resistance of plant. In *9. International Plant Biomechanics Conference. PlantBiomech 2018*, 2018.

Christophe Genolini, Xavier Alacoque, Mariane Sentenac, and Catherine Arnaud. kml and kml3d: R packages to cluster longitudinal data. *Journal of Statistical Software*, 65:1–34, 2015.

Ritesh Ghosh, Ratnesh Chandra Mishra, Bosung Choi, Young Sang Kwon, Dong Won Bae, Soo-Chul Park, Mi-Jeong Jeong, and Hanhong Bae. Exposure to sound vibrations lead to transcriptomic, proteomic and hormonal changes in arabidopsis. *Scientific reports*, 6(1):1–17, 2016.

Ritesh Ghosh, Mayank A Gururani, Lakshmi N Ponpandian, Ratnesh C Mishra, Soo-Chul Park, Mi-Jeong Jeong, and Hanhong Bae. Expression analysis of sound vibration-regulated genes by touch treatment in arabidopsis. *Frontiers in plant science*, 8:100, 2017.

Ritesh Ghosh, Adelin Barbacci, and Nathalie Leblanc-Fournier. Mechanostimulation: a promising alternative for sustainable agriculture practices. *Journal of Experimental Botany*, 72(8):2877–2888, 2021.

Ritesh Ghosh, Juliette Roue, Jerome Franchel, Amit Paul, and Nathalie Leblanc-Fournier. Temporal modification of h3k9/14ac and h3k4me3 histone marks mediates mechano-responsive gene expression during the accommodation process in poplar. *bioRxiv*, pages 2023–02, 2023a.

Ritesh Ghosh, Juliette Roué, Jérôme Franchel, Amit Paul, and Nathalie Leblanc-Fournier. Temporal modification of H3K9/14ac and H3K4me3 histone marks mediates mechano-responsive gene expression during the accommodation process in poplar. *bioRxiv*, page 2023.02.12.526104, February 2023b. doi: 10.1101/2023.02.12.526104. URL https://doi.org/10.1101/2023.02.12.526104. Publisher: Cold Spring Harbor Laboratory.

Kwang-Il Goh, Michael E Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-László Barabási. The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685–8690, 2007.

Albert Goldbeter. A model for circadian oscillations in the drosophila period protein (per). *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 261(1362):319–324, 1995.

Reinhard Guthke, Ulrich Möller, Martin Hoffmann, Frank Thies, and Susanne Töpfer. Dynamic network reconstruction from gene expression data applied to immune response during bacterial infection. *Bioinformatics*, 21(8):1626–1634, 2005.

Olivier Hamant and Elizabeth S Haswell. Life behind the wall: Sensing mechanical cues in plants, 2017. ISSN 17417007.

Reda Hassanien Emam Hassanien and Bao Ming LI. Dual effect of audible sound technology on the growth and endogenous hormones of strawberry. *Agricultural Engineering International: CIGR Journal*, 22(3):262–273, 2020.

Trevor Hastie and Junyang Qian. Glmnet vignette. *Retrieved June*, 9(2016):1–30, 2014.

M T Hauser, B Harr, and C Schlötterer. Trichome distribution in arabidopsis thaliana and its close relative arabidopsis lyrata: molecular analysis of the candidate gene glabrous1. *Molecular biology and evolution*, 18:1754–1763, 2001. ISSN 0737-4038. doi: 10.1093/oxfordjournals.molbev.a003963.

Nicholas A Heard, Christopher C Holmes, David A Stephens, David J Hand, and George Dimopoulos. Bayesian coclustering of anopheles gene expression time series: study of immune defense response to multiple experimental challenges. *Proceedings of the National Academy of Sciences*, 102(47):16939–16944, 2005.

Michael Hecker, Sandro Lambeck, Susanne Toepfer, Eugene Van Someren, and Reinhard Guthke. Gene regulatory network inference: data integration in dynamic models—a review. *Biosystems*, 96(1):86–103, 2009.

Reinhart Heinrich and Stefan Schuster. *The regulation of cellular systems*. Springer Science & Business Media, 2012.

Yosef Hochberg and Yoav Benjamini. More powerful procedures for multiple significance testing. *Statistics in medicine*, 9(7):811–818, 1990.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.

Huei-Chung Huang, Yi Niu, and Li-Xuan Qin. Differential expression analysis for rna-seq: an overview of statistical methods and computational software: supplementary issue: sequencing platform modeling and analysis. *Cancer informatics*, 14: CIN–S21631, 2015.

François Husson, Sébastien Lê, and Jérôme Pagès. *Exploratory multivariate analysis by example using R*. CRC press, 2017.

Koji Iwayama, Yuri Aisaka, Natsumaro Kutsuna, and Atsushi J Nagano. Fit: statistical modeling tool for transcriptome dynamics under fluctuating field conditions. *Bioinformatics*, 33(11):1672–1680, 2017.

M. J. Jaffe. Thigmomorphogenesis: The response of plant growth and development to mechanical stimulation. *Planta*, 114(2):143–157, 1973. ISSN 0032-0935. doi: 10.1007/BF00387472. URL http://link.springer.com/10.1007/BF00387472.

Mi-Jeong Jeong, Chang-Ki Shim, Jin-Ohk Lee, Hawk-Bin Kwon, Yang-Han Kim, Seong-Kon Lee, Myeong-Ok Byun, and Soo-Chul Park. Plant gene responses to frequency-specific sound signals. *Molecular breeding*, 21(2):217–226, 2008.

Mi-Jeong Jeong, Jung-Il Cho, Sung-Han Park, Kyung-hwan Kim, Seong Kon Lee, Taek-Ryoun Kwon, Soo-Chul Park, Zamin Shaheed Siddiqui, et al. Sound frequencies induce drought tolerance in rice plant. *Pak J Bot*, 46:2015–2020, 2014.

Jonathan DG Jones and Jeffery L Dangl. The plant immune system. *nature*, 444 (7117):323–329, 2006.

Jihye Jung, Seon-Kyu Kim, Joo Y Kim, Mi-Jeong Jeong, and Choong-Min Ryu. Beyond chemical triggers: evidence for sound-evoked physiological reactions in plants. *Frontiers in plant science*, 9:25, 2018.

Itzhak Khait, Ohad Lewin-Epstein, Raz Sharon, Kfir Saban, Revital Goldstein, Yehuda Anikster, Yarden Zeron, Chen Agassy, Shaked Nizan, Gayl Sharabi, et al. Sounds emitted by plants under stress are airborne and informative. *Cell*, 186(7): 1328–1336, 2023.

Joo-Yeol Kim, Jin-Su Lee, Taek-Ryoun Kwon, Soo-In Lee, Jin-A Kim, Gyu-Myoung Lee, Soo-Chul Park, and Mi-Jeong Jeong. Sound waves delay tomato fruit ripening by negatively regulating ethylene biosynthesis and signaling genes. *Postharvest Biology and Technology*, 110:43–50, 2015.

Joo Yeol Kim, Seon-Kyu Kim, Jihye Jung, Mi-Jeong Jeong, and Choong-Min Ryu. Exploring the sound-modulated delay in tomato ripening through expression analysis of coding and non-coding rnas. *Annals of botany*, 122(7):1231–1244, 2018.

Joo Yeol Kim, Ye Eun Kang, Soo In Lee, Jin A Kim, Muthusamy Muthusamy, and Mi-Jeong Jeong. Sound waves affect the total flavonoid contents in medicago sativa, brassica oleracea and raphanus sativus sprouts. *Journal of the Science of Food and Agriculture*, 100(1):431–440, 2020.

Shuhei Kimura, Kaori Ide, Aiko Kashihara, Makoto Kano, Mariko Hatakeyama, Ryoji Masui, Noriko Nakagawa, Shigeyuki Yokoyama, Seiki Kuramitsu, and Akihiko Konagaya. Inference of s-system models of genetic networks using a cooperative coevolutionary algorithm. *Bioinformatics*, 21(7):1154–1163, 2005.

Young Sang Kwon, Mi-Jeong Jeong, Jaeyul Cha, Sung Woo Jeong, Soo-Chul Park, Sung Chul Shin, Woo Sik Chung, Hanhong Bae, and Dong-Won Bae. Comparative proteomic analysis of plant responses to sound waves in arabidopsis. *Journal of Plant Biotechnology*, 39(4):261–272, 2012.

Jörn Lämke and Isabel Bäurle. Epigenetic and chromatin-based mechanisms in environmental stress adaptation and stress memory in plants. *Genome biology*, 18 (1):1–11, 2017.

David S Latchman. Transcription factors: an overview. *The international journal of biochemistry & cell biology*, 29(12):1305–1312, 1997.

Steffen L Lauritzen. *Graphical models*, volume 17. Clarendon Press, 1996.

Charity W Law, Yunshun Chen, Wei Shi, and Gordon K Smyth. voom: Precision weights unlock linear model analysis tools for rna-seq read counts. *Genome biology*, 15(2):1–17, 2014.

Charity W Law, Monther Alhamdoosh, Shian Su, Gordon K Smyth, and Matthew E Ritchie. Rna-seq analysis is easy as 1-2-3 with limma, glimma and edger. *F1000Research*, 5, 2016.

Dennis Lee, Diana H. Polisensky, and Janet Braam. Genome-wide identification of touch- and darkness-regulated Arabidopsis genes: A focus on calmodulin-like and XTH genes. *New Phytologist*, 165(2):429–444, 2005. ISSN 0028646X. doi: 10.1111/j.1469-8137.2004.01238.x. ISBN: 0028-646X (Print)\n0028-646X (Linking).

Ophélie Léger, Frédérick Garcia, Mehdi Khafif, Sebastien Carrere, Nathalie Leblanc-Fournier, Aroune Duclos, Vincent Tournat, Eric Badel, Marie Didelon, Aurélie Le Ru, et al. Pathogen-derived mechanical cues potentiate the spatio-temporal implementation of plant defense. *BMC biology*, 20(1):292, 2022.

Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and samtools. *bioinformatics*, 25(16):2078–2079, 2009.

Alexandre Liapounoff. Problème général de la stabilité du mouvement. In *Annales de la Faculté des sciences de Toulouse: Mathématiques*, volume 9, pages 203–474, 1907.

Jason Lines and Anthony Bagnall. Time series classification with ensembles of elastic distance measures. *Data Mining and Knowledge Discovery*, 29:565–592, 2015.

Yu Ling, Natalia Serrano, Ge Gao, Mohamed Atia, Morad Mokhtar, Yong H Woo, Jeremie Bazin, Alaguraj Veluchamy, Moussa Benhamed, Martin Crespi, et al. Thermopriming triggers splicing memory in arabidopsis. *Journal of Experimental Botany*, 69(10):2659–2675, 2018.

Haipei Liu, Amanda J Able, and Jason A Able. Priming crops for the future: rewiring stress memory. *Trends in plant science*, 27(7):699–716, 2022.

S B Liu and J J Jiao. Arabidopsis leaf trichomes as acoustic antennae. *Biophysj*, 113:1–5, 2017. ISSN 00063495. doi: 10.1016/j.bpj.2017.07.035. URL https://doi.org/10.1016/j.bpj.2017.07.035.

Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):1–21, 2014.

Ophélie Léger, Frédérick Garcia, Mehdi Khafif, Sebastien Carrere, Nathalie Leblanc-Fournier, Aroune Duclos, Vincent Tournat, Eric Badel, Marie Didelon, Aurélie Le Ru, Sylvain Raffaele, and Adelin Barbacci. Pathogen-derived mechanical cues potentiate the spatio-temporal implementation of plant defense. *BMC Biology*, 20:292, 12 2022. ISSN 1741-7007. doi: 10.1186/s12915-022-01495-w. URL https://bmcbiol.biomedcentral.com/articles/10.1186/s12915-022-01495-w.

Dimitrije Markovic, Ilaria Colzi, Cosimo Taiti, Swayamjit Ray, Romain Scalone, Jared Gregory Ali, Stefano Mancuso, and Velemir Ninkovic. Airborne signals synchronize the defenses of neighboring plants in response to touch. *Journal of experimental botany*, 70(2):691–700, 2019.

Ludovic Martin, Nathalie Leblanc-Fournier, Jean Louis Julien, Bruno Moulia, and Catherine Coutand. Acclimation kinetics of physiological and molecular responses of plants to multiple mechanical loadings. *Journal of Experimental Botany*, 61: 2403–2412, 2010a. ISSN 00220957. doi: 10.1093/jxb/erq069.

Ludovic Martin, Nathalie Leblanc-Fournier, Jean-Louis Julien, Bruno Moulia, and Catherine Coutand. Acclimation kinetics of physiological and molecular responses

of plants to multiple mechanical loadings. *Journal of Experimental Botany*, 61(9): 2403–2412, 2010b.

Ludovic Martin, Nathalie Leblanc-Fournier, Jean Louis Julien, Bruno Moulia, and Catherine Coutand. Acclimation kinetics of physiological and molecular responses of plants to multiple mechanical loadings. *Journal of Experimental Botany*, 61 (9):2403–2412, 2010c. ISSN 00220957. doi: 10.1093/jxb/erq069. ISBN: 1460-2431 (Electronic)\r0022-0957 (Linking).

Shawn Martin, Zhaoduo Zhang, Anthony Martino, and Jean-Loup Faulon. Boolean dynamics of genetic regulatory networks inferred from microarray time series data. *Bioinformatics*, 23(7):866–874, 2007.

Ainhoa Martinez-Medina, Victor Flors, Martin Heil, Brigitte Mauch-Mani, Corné MJ Pieterse, Maria J Pozo, Jurriaan Ton, Nicole M van Dam, and Uwe Conrath. Recognizing plant defense priming. *Trends in plant science*, 21(10): 818–822, 2016.

Hirotaka Matsumoto, Hisanori Kiryu, Chikara Furusawa, Minoru SH Ko, Shigeru BH Ko, Norio Gouda, Tetsutaro Hayashi, and Itoshi Nikaido. Scode: an efficient regulatory network inference algorithm from single-cell rna-seq during differentiation. *Bioinformatics*, 33(15):2314–2321, 2017.

Mamoru Matsumura, Mika Nomoto, Tomotaka Itaya, Yuri Aratani, Mizuki Iwamoto, Takakazu Matsuura, Yuki Hayashi, Tsuyoshi Mori, Michael J. Skelly, Yoshiharu Y. Yamamoto, Toshinori Kinoshita, Izumi C. Mori, Takamasa Suzuki, Shigeyuki Betsuyaku, Steven H. Spoel, Masatsugu Toyota, and Yasuomi Tada. Mechanosensory trichome cells evoke a mechanical stimuli–induced immune response in arabidopsis thaliana. *Nature Communications*, 13:1–15, 12 2022. ISSN 20411723. doi: 10.1038/s41467-022-28813-8. URL https://doi.org/10.1038/s41467-022-28813-8.

RP McCall. Sound, speech and hearing. *Physics of the Human Body*, 116, 2010.

Sunnie Grace McCalla, Alireza Fotuhi Siahpirani, Jiaxin Li, Saptarshi Pyne, Matthew Stone, Viswesh Periyasamy, Junha Shin, and Sushmita Roy. Identifying strengths and weaknesses of methods for computational network inference from single-cell rna-seq data. *G3: Genes, Genomes, Genetics*, 13(3):jkad004, 2023.

Ratnesh Chandra Mishra and Hanhong Bae. Plant cognition: ability to perceive 'touch'and 'sound'. *Sensory Biology of Plants*, pages 137–162, 2019.

Ratnesh Chandra Mishra, Ritesh Ghosh, and Hanhong Bae. Plant acoustics: in the search of a sound mechanism for sound signaling in plants. *Journal of experimental botany*, 67(15):4483–4494, 2016.

Michel Morange. Les mirages de l'épigénétique. *Critique*, (2):153–158, 2011.

Thomas Morgan. Epigénétique: quand l'environnement marque nos gènes.

Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature methods*, 5(7):621–628, 2008.

Bruno Moulia, Stéphane Douady, and Olivier Hamant. Fluctuations shape plants through proprioception. *Science*, 372(6540):eabc6868, April 2021. ISSN 0036-8075. doi: 10.1126/science.abc6868. URL https://www.sciencemag.org/lookup/doi/10.1126/science.abc6868. Publisher: American Association for the Advancement of Science.

Vito MR Muggeo et al. Segmented: an r package to fit regression models with broken-line relationships. *R news*, 8(1):20–25, 2008.

Kary B Mullis and Fred A Faloona. [21] specific synthesis of dna in vitro via a polymerase-catalyzed chain reaction. In *Methods in enzymology*, volume 155, pages 335–350. Elsevier, 1987.

Christopher C Mundt. Pyramiding for resistance durability: theory and practice. *Phytopathology*, 108(7):792–802, 2018.

Christopher A Penfold and David L Wild. How to infer gene networks from expression profiles, revisited. *Interface focus*, 1(6):857–870, 2011.

Rémi Peyraud, Malick Mbengue, Adelin Barbacci, and Sylvain Raffaele. Intercellular cooperation in a fungal plant pathogen facilitates host colonization. *Proceedings of the National Academy of Sciences of the United States of America*, page 201811267, February 2019. ISSN 1091-6490. doi: 10.1073/pnas.1811267116. URL http://www.ncbi.nlm.nih.gov/pubmed/30728304. Publisher: National Academy of Sciences.

Remi Peyronnet, Daniel Tran, Tiffanie Girault, and Jean-Marie Frachisse. Mechanosensitive channels: feeling tension in a world under pressure. *Frontiers in plant science*, 5:558, 2014a.

Rémi Peyronnet, Daniel Tran, Tiffanie Girault, and Jean-Marie Frachisse. Mechanosensitive channels: feeling tension in a world under pressure. *Frontiers in*

plant science, 5(October):558, 2014b. ISSN 1664-462X. doi: 10.3389/fpls.2014.00558. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4204436&tool=pmcentrez&rendertype=abstract. ISBN: 1664-462X.

Marie-Laure Pilet-Nayel, Benoît Moury, Valérie Caffier, Josselin Montarry, Marie-Claire Kerlan, Sylvain Fournet, Charles-Eric Durel, and Régine Delourme. Quantitative resistance to plant pathogens in pyramiding strategies for durable crop protection. *Frontiers in plant science*, 8:1838, 2017.

Jesse A Poland, Peter J Balint-Kurti, Randall J Wisser, Richard C Pratt, and Rebecca J Nelson. Shades of gray: the world of quantitative disease resistance. *Trends in plant science*, 14(1):21–29, 2009.

Lise Pomiès, Mélanie Decourteix, Jérôme Franchel, Bruno Moulia, and Nathalie Leblanc-Fournier. Poplar stem transcriptome is massively remodelled in response to single or repeated mechanical stimuli. *BMC genomics*, 18:1–16, 2017.

Lirong Qi, Guanghui Teng, Tianzhen Hou, Baoying Zhu, and Xiaona Liu. Influence of sound wave stimulation on the growth of strawberry in sunlight greenhouse. In *Computer and Computing Technologies in Agriculture III: Third IFIP TC 12 International Conference, CCTA 2009, Beijing, China, October 14-17, 2009, Revised Selected Papers 3*, pages 449–454. Springer, 2010.

Yu-Chuan Qin, Won-Chu Lee, Young-Cheol Choi, and Tae-Wan Kim. Biochemical and physiological changes in plants as a result of different sonic exposures. *Ultrasonics*, 41(5):407–411, 2003.

Jean-Pierre Richard and M Ksouri. Systèmes à retard. *chapitre*, 6:235–288, 2008.

Matthew E Ritchie, Belinda Phipson, DI Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*, 43(7):e47–e47, 2015.

Mark D Robinson and Alicia Oshlack. A scaling normalization method for differential expression analysis of rna-seq data. *Genome biology*, 11(3):1–9, 2010.

Ana Rodrigo-Moreno, Nadia Bazihizina, Elisa Azzarello, Elisa Masi, Daniel Tran, François Bouteau, Frantisek Baluska, and Stefano Mancuso. Root phonotropism: early signalling events following sound perception in arabidopsis roots. *Plant Science*, 264:9–15, 2017.

Juan José Rodríguez and Carlos J Alonso. Support vector machines of interval-based features for time series classification. In *International Conference on Innovative*

Techniques and Applications of Artificial Intelligence, pages 244–257. Springer, 2004.

Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

Fabrice Roux, Derry Voisin, Thomas Badet, Claudine Balagué, Xavier Barlet, Carine Huard-Chauveau, Dominique Roby, and Sylvain Raffaele. Resistance to phytopathogens e tutti quanti: placing plant quantitative disease resistance on the map. *Molecular plant pathology*, 15(5):427–432, 2014.

Karen Sachs, Omar Perez, Dana Pe'er, Douglas A Lauffenburger, and Garry P Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529, 2005.

Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing*, 26(1):43–49, 1978.

Arun Sampathkumar, Pawel Krupinski, Raymond Wightman, Pascale Milani, Alexandre Berquand, Arezki Boudaoud, Olivier Hamant, Henrik Jönsson, and Elliot M Meyerowitz. Subcellular and supracellular mechanical stress prescribes cytoskeleton behavior in \textit{Arabidopsis} cotyledon pavement cells. *eLife*, 3: e01967, 2014. ISSN 2050-084X. doi: 10.7554/eLife.01967. URL https://doi.org/10.7554/eLife.01967. Publisher: eLife Sciences Publications, Ltd.

Guido Sanguinetti et al. Gene regulatory network inference: an introductory survey. In *Gene Regulatory Networks*, pages 1–23. Springer, 2019.

Alexis Sarda-Espinosa, Maintainer Alexis Sarda, and TRUE LazyData. Package 'dtwclust'. *Pobrane z: http://cran. ma. imperial. ac. uk/web/packages/dtwclust/dtwclust. pdf*, 2018.

Serge Savary, Laetitia Willocquet, Sarah Jane Pethybridge, Paul Esker, Neil McRoberts, and Andy Nelson. The global burden of pathogens and pests on major food crops. *Nature ecology & evolution*, 3(3):430–439, 2019.

Gert Sclep, Joke Allemeersch, Robin Liechti, Björn De Meyer, Jim Beynon, Rishikesh Bhalerao, Yves Moreau, Wilfried Nietfeld, Jean-Pierre Renou, Philippe Reymond, et al. Catma, a comprehensive genome-scale resource for silencing and transcript profiling of arabidopsis genes. *BMC bioinformatics*, 8(1):1–13, 2007.

Andrew J Severin, Jenna L Woody, Yung-Tsi Bolon, Bindu Joseph, Brian W Diers, Andrew D Farmer, Gary J Muehlbauer, Rex T Nelson, David Grant, James E Specht, et al. Rna-seq atlas of glycine max: a guide to the soybean transcriptome. *BMC plant biology*, 10:1–16, 2010.

J Shipman, Jerry D Wilson, and CA Higgins. Waves and sound. *An Introduction to Physical Science*, pages 134–142, 2012.

Chao Sima, Jianping Hua, and Sungwon Jung. Inference of gene regulatory networks using time-series data: a survey. *Current genomics*, 10(6):416–429, 2009.

Prashant Singh, Shweta Yekondi, Po-Wen Chen, Chia-Hong Tsai, Chun-Wei Yu, Keqiang Wu, and Laurent Zimmerli. Environmental history modulates arabidopsis pattern-triggered immunity in a histone acetyltransferase1–dependent manner. *The Plant Cell*, 26(6):2676–2688, 2014.

Alicia T Specht and Jun Li. Leap: constructing gene co-expression networks for single-cell rna-sequencing data using pseudotime ordering. *Bioinformatics*, 33(5): 764–766, 2017.

Justine Sucher, Malick Mbengue, Axel Dresen, Marielle Barascud, Marie Didelon, Adelin Barbacci, and Sylvain Raffaele. Phylotranscriptomics of the pentapetalae reveals frequent regulatory variation in plant local responses to the fungal pathogen sclerotinia sclerotiorum. *The Plant Cell*, page tpc.00806.2019, 4 2020. ISSN 1040-4651. doi: 10.1105/tpc.19.00806. URL http://www.plantcell.org/lookup/doi/10.1105/tpc.19.00806.

Nozomu Takahashi, Yoshito Hirata, Kazuyuki Aihara, and Paloma Mas. A Hierarchical Multi-oscillator Network Orchestrates the Arabidopsis Circadian System. *Cell*, 163(1):148–159, September 2015. ISSN 0092-8674. doi: 10.1016/j.cell.2015.08.062. URL https://www.sciencedirect.com/science/article/pii/S0092867415011149.

Yi Tao, Zhiyi Xie, Wenqiong Chen, Jane Glazebrook, Hur-Song Chang, Bin Han, Tong Zhu, Guangzhou Zou, and Fumiaki Katagiri. Quantitative nature of arabidopsis responses during compatible and incompatible interactions with the bacterial pathogen pseudomonas syringae. *The Plant Cell*, 15(2):317–330, 2003.

R Core Team, Maintainer R Core Team, MASS Suggests, and S Matrix. Package "stats.". *The R Stats Package*, 2018.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Rodrigo Hernán Tomas-Grau, Fernando José Requena-Serra, Verónica Hael-Conrad, Martín Gustavo Martínez-Zamora, María Fernanda Guerrero-Molina, and Juan Carlos Díaz-Ricci. Soft mechanical stimulation induces a defense response against botrytis cinerea in s trawberry. *Plant cell reports*, 37:239–250, 2018.

Daniel Tran, Tiffanie Girault, Marjorie Guichard, Sébastien Thomine, Nathalie Leblanc-Fournier, Bruno Moulia, Emmanuel de Langre, Jean Marc Allain, and Jean Marie Frachisse. Cellular transduction of mechanical oscillations in plants by the plasma-membrane mechanosensitive channel MSL10. *Proceedings of the National Academy of Sciences of the United States of America*, 118(1), January 2021. ISSN 10916490. doi: 10.1073/PNAS.1919402118. URL https://www.pnas.org/content/118/1/e1919402118. Publisher: National Academy of Sciences.

Likun Wang, Zhixing Feng, Xi Wang, Xiaowo Wang, and Xuegong Zhang. Degseq: an r package for identifying differentially expressed genes from rna-seq data. *Bioinformatics*, 26(1):136–138, 2010.

Oliver Windram, Priyadharshini Madhou, Stuart McHattie, Claire Hill, Richard Hickman, Emma Cooke, Dafyd J Jenkins, Christopher A Penfold, Laura Baxter, Emily Breeze, et al. Arabidopsis defense against botrytis cinerea: chronology and regulation deciphered by high-resolution temporal transcriptomic analysis. *The Plant Cell*, 24(9):3530–3557, 2012.

Daniela M Witten. Classification and clustering of sequencing data using a poisson model. 2011.

Cheng-Wei Yao, Ban-Dar Hsu, and Bor-Sen Chen. Constructing gene regulatory networks for long term photosynthetic light acclimation in arabidopsis thaliana. *BMC bioinformatics*, 12(1):1–16, 2011.

Xinyou Yin, Paul C Struik, Junfei Gu, and Huaqi Wang. Modelling qtl-trait-crop relationships: past experiences and future prospects. *Crop systems biology: narrowing the gaps between crop modelling and genetics*, pages 193–218, 2016.

Neil A Youngson and Emma Whitelaw. Transgenerational epigenetic effects. *Annu. Rev. Genomics Hum. Genet.*, 9:233–257, 2008.

11 Annexe 1 : Organigramme de l'analyse des données RNA-seq du son.

La figure 11.1 résume les différentes étapes de l'analyse des données d'expression de gènes présenté dans ce chapitre au cours de la thèse : de l'analyse différentielles des données transcriptomiques mesurées avec la technique RNA-Seq jusqu'à l'analyse intégrative de ces données avec des données phénotypique.

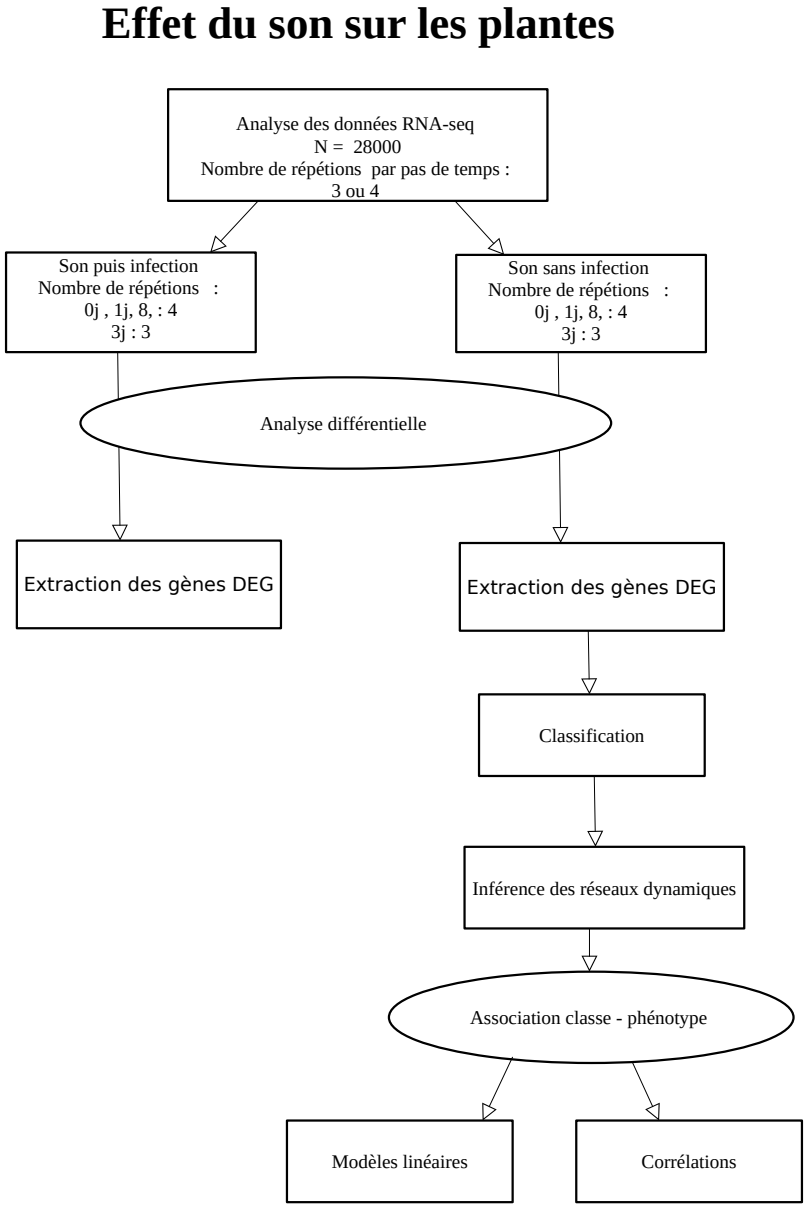


Figure 11.1: Organigramme de l'analyse des données RNA-seq du son. Pour l'analyse différentielle des tests deux à deux ont été testés. Puis, nous avons sélectionnés les gènes différentiellement exprimés (DEG) pour appliquer une méthode de classification temporelle. À partir des résultats de classification nous avons inférer des réseaux dynamiques entres classes des gènes pour enfin associé ce réseau aux expressions des phénotypes obtenus.

12 Annexe 2 : Défense d'*Arabidopsis thaliana* contre *Botrytis cinerea*

Arabidopsis thaliana présente une résistance quantitative aux maladies face aux champignons nécrotrophes tels que *B. cinerea* et *S. sclerotinia*. Le QDR végétal est multigénique et implique un grand nombre de gènes (plus de 10'000 gènes sont modulés lors de l'infection). D'un point de vue biologique, ce travail vise à révéler la structure QDR. D'un point de vue modéliste, ce travail est l'occasion de tirer profit des données publiées [Windram et al., 2012] pour développer un pipeline de méthodes allant du clustering à l'inférence de réseau dynamique en utilisant des données temporelles régulièrement espacées.

12.0.1 Expression de gènes

Nous avons utilisé les données GSE29642 fournies par Windram et al (Windram et al.,2012). Des séries temporelles d'expression des gènes ont été générées à partir de feuilles détachées d'*Arabidopsis thaliana* âgées de 4 semaines infectées par le pathogène nécrotrophe *Botrytis cinerea*. Les expressions des gènes ont été mesurées pour les feuilles infectées et moquées (contrôle) toutes les 2 heures pendant une durée de 48 heures (Figure 12.1) (24 points temporels séparés de 2 h ; quatre répétitions biologiques et une moyenne de trois répétitions techniques pour chaque point temporel dans chaque condition). L'analyse de l'expression a été réalisée à l'aide de puces CATMA (un micro-réseau complet de transcriptome d'*Arabidopsis thaliana*) (Sclep et al.,2007), d'ADNc de feuilles simples et d'une conception statistique en boucle d'hybridations, conduisant à une résolution élevée et hautement répliquée. (Windram et al.,2012).

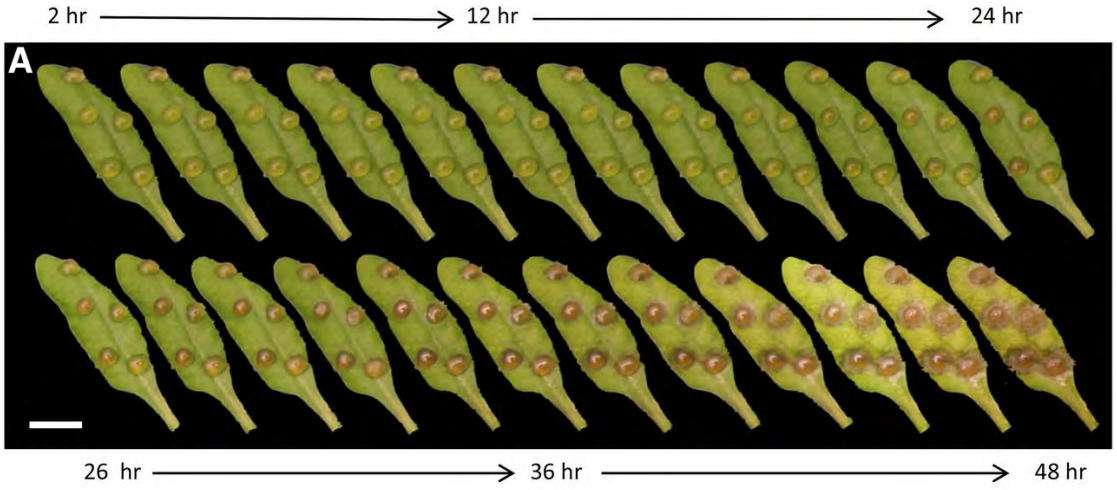


Figure 12.1: Infection temporelle de *Botrytis cineria* sur les feuilles d'*Arabidopsis thaliana* (Windram et al.,2012).

12.1 Méthode de classification hiérarchique temporelle

Lorsqu'on travaille avec des série temporelles long et régulièrement espacées, des méthodes de classification assez compliqué seront visé. Le but de cette section est d'utiliser une méthode de classification simple et adaptée à ces types des données.

L'idée principale est d'utiliser la **méthode de classification hiérarchique sur les composants principaux** - ACP pour classer les trajectoires des gènes. La classification avec un ACP a été appliquée sur K intervalles temporels (K déterminé à l'aide d'un ACP [Husson et al., 2017]). Pour chaque intervalle de temps, une classification hiérarchique sur les composantes principales a été appliquée. Cet algorithme est composé de deux étapes ; L'analyse en composantes principales (ACP) a été utilisée pour réduire la dimension des données en quelques variables contenant les informations les plus importantes dans les données dans chaque intervalle. Puis une classification hiérarchique a été effectuée sur les résultats de l'ACP. L'étape de l'ACP peut être considérée comme une étape qui réduit le bruit dans les données, ce qui peut conduire à une classification plus stable.
Pour la classification hiérarchique temporelle, on va appliquer la méthode de classification sur les composants principaux sur le résultat de classification de l'intervalle précédent. Donc on a deux étapes à faire avant de commencer la classification ; on va découper l'intervalle temporel de départ en K sous intervalles et on va fixer le nombre des classes C pour chaque intervalle. Les deux principales

questions posées pour cette méthode sont alors : À quel instant on coupe l'intervalle ? et combien de classe on va travailler avec ?
Pour choisir les instants qu'on va les utiliser pour chaque sous intervalle on applique une ACP sur l'intervalle de temps de départ. On utilisant les résultats de l'ACP, on va choisir les instants qui sont assez proches du premier axe et qui sont assez loin de l'origine. C'est sont les instants le plus corrélés avec le premier axe. La dernière étape qui consiste à fixer le nombre de classe est faite on utilisant l'indice de Davies-Bouldin : la moyenne du rapport maximal entre la distance d'un point au centre de son groupe et la distance entre deux centres de groupes.

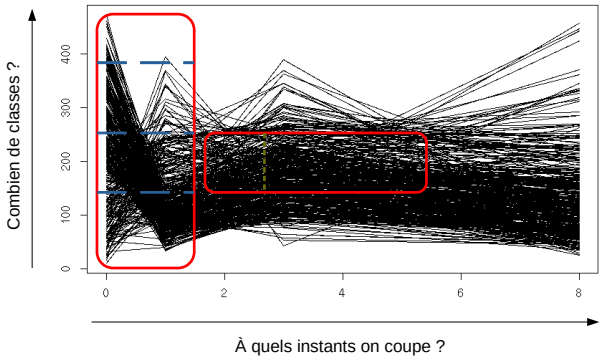


Figure 12.2: Classification hiérarchique temporelle

Algorithm 1 hierarchical classification

input: Data: A table of T colums and N variables , N > >T.
output: Clusters that groups genes patterns .
res.pca = PCA(data, scale.unit = FALSE, graph = FALSE)

let delta1 be the first data part
res1.pca = PCA(delta1,scale.unit = FALSE, graph = FALSE)
hc1 = HCPC(res1.pca, nb.clust=-1, graph=FALSE)

1: *let l_1 be the levels of the clusters of hc1*
2: **for** for nc in 1 to length(l_1) **do**
3: *let delta2 be the second data part in which the first clustering level is nc*
4: res2.pca <- PCA(delta2, scale.unit = FALSE, graph = FALSE)
5: hc2 <- HCPC(res2.pca, nb.clust=-1, graph=FALSE)
6: *let l_2 be the levels of the clusters of hc2*
7: **for** kc in 1to length(l_2) **do**
8: *let delta3 be the third data part in which the first clustering level is nc and the second clustering level is kc*
9: res3.pca = PCA(delta3, scale.unit = FALSE, graph = FALSE)
10: hc3 = HCPC(res3.pca, nb.clust=-1, graph=FALSE)
11: **end for**
12: **end for**

Les gènes modulés ont été disposés dans des classes à l'aide de la méthode de classification hiérarchique temporelle. L'algorithme de classification a été utilisé sur 3 plages temporelles (t= 2 à t=22 , t= 24 à t=38 et t= 40 à t=48). Pour chaque plage de temps, une classification hiérarchique sur les composantes principales a été appliquée. Le nombre des classes étaient fixées en utilisant l'indice de silhouette. À l'issue de cette étape 27 classes ont été formées et qui vont être utilisé pour l'inférence de réseau.

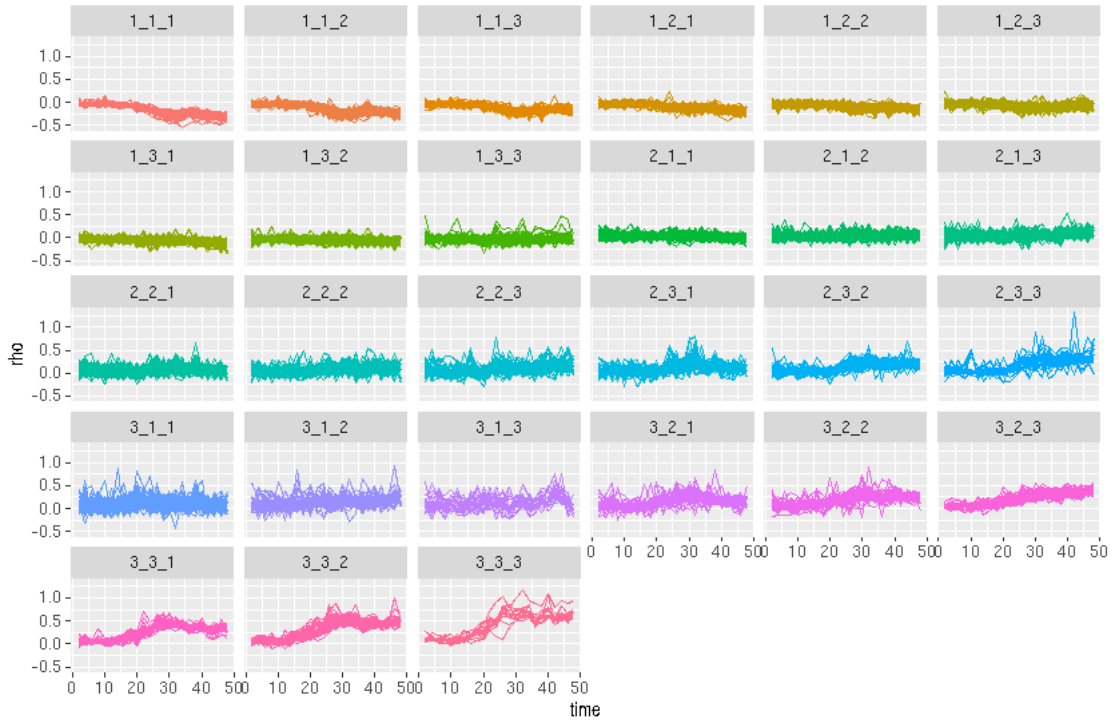


Figure 12.3: Résultat de classification hiérarchique temporelles des expressions des gènes différentiellement exprimés.

Inférence de réseau

La classification a permis de regrouper les gènes en 27 classes différents respectant la forme de chaque expression de gènes. Pour qu'on puisse modéliser (inférer) ces expressions, une valeur représentante de chaque classe sera sélectionné ; cette expression peut être un gène spécifique connu (là on va utiliser des connaissance biologique à priori) , l'expression moyenne, les quantiles ou même les expressions les plus proches de l'expression moyenne de chaque classe. Dans notre étude, et puisque nous avons classifié les données sans des connaissances biologiques à priori, nous avons décidé d'utiliser l'expression moyenne de chaque classe comme représentant de chaque classe et qui va servir pour inférer le réseau entre les différentes classes. Pour décrire la dynamique de ces expressions moyennes, on a utilisé les modèles autorégressif d'ordre 1 :

$$X_t = AX_{t-1} + \epsilon_t \tag{12.1}$$

Avec X_t sera le vecteur des expressions moyennes de chaque classe au temps t et A sera la matrice des coefficients qui sera estimé à l'aide d'une méthode d'estimation pénalisé (lasso).

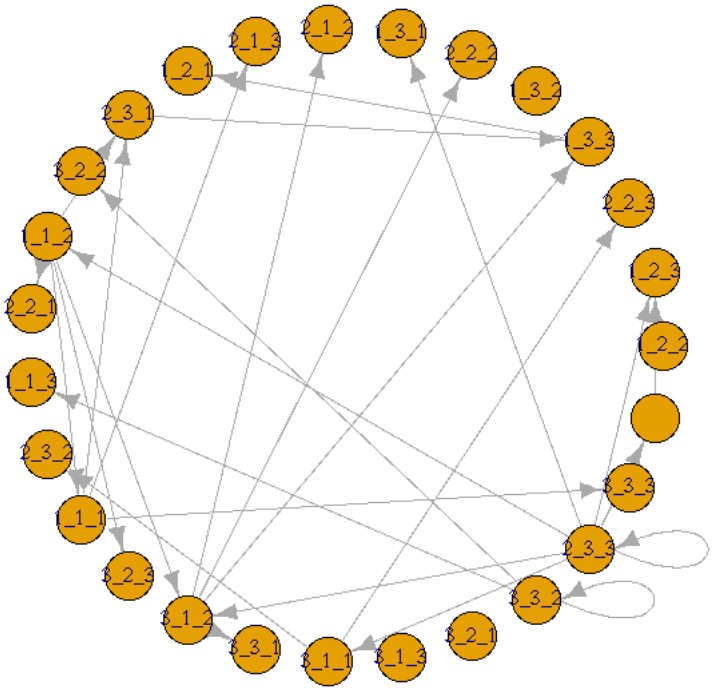


Figure 12.4: Réseau obtenu en utilisant le modèle autorégressif d'ordre 1.

12.1.1 Retour à la biologies

L'interprétation des résultats biologiquement repose sur deux types d'analyse : d'une part, regarder si le réseau inféré est spécifique au QDR de *B cinerea* et d'autre part l'analyse enrichissement issu de l'ontologie GO (Gene Ontology) des gènes Le deux type d'analyse sert à donner un sens biologique aux résultats statistiques trouvées.

Spécificité de réseau inféré:

Pour tester si le réseau inféré est spécifique au QDR de *B cinerea*, nous avons vérifié si des gènes en clusters étaient également exprimés lors d'infections de 4 autres champignons nécrotrophes tels que *Sclerotinia sclerotiorum, Verticillium et Alternaria*. Nous avons trouvé que les gènes modulés pendant l'infection par *B. cinerea* étaient majoritairement modulés lors de l'infection par d'autres champignons (Figure 12.5) et non spécifiques du QDR à *B cinerea*.

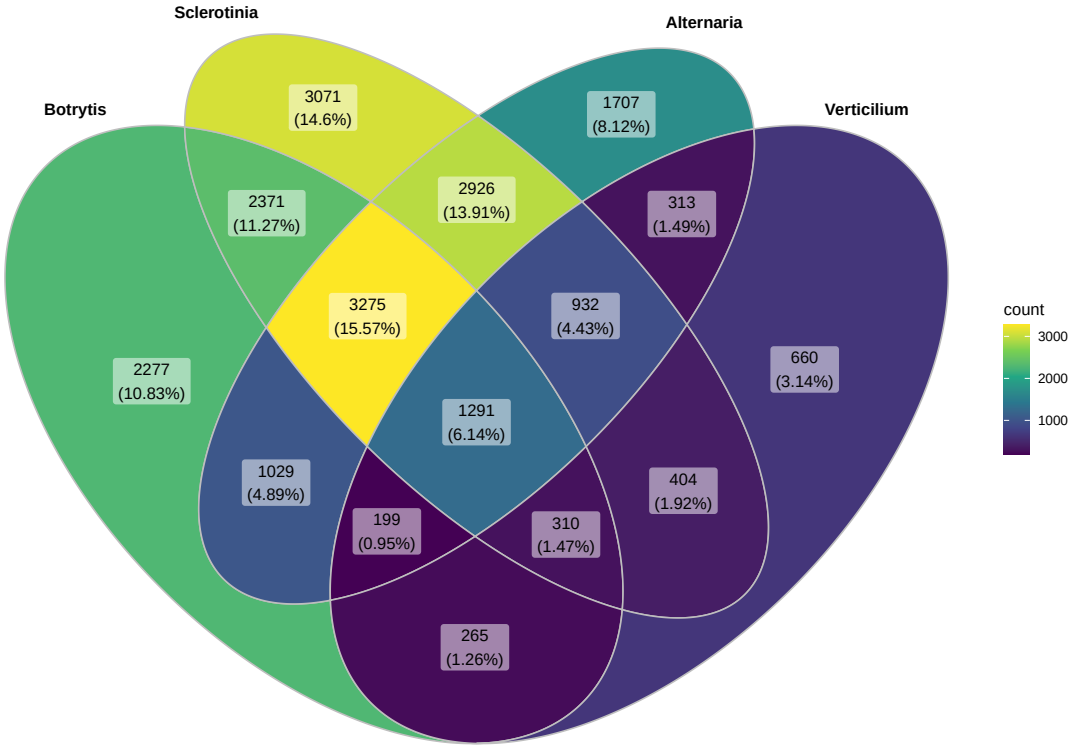


Figure 12.5: Diagramme de Venn des gènes différentiellement exprimés lors de l'infection à *B. cinerea, Sclerotinian, Alternaria and Verticillum*

Analyse d'enrichissement:
Alexa et al. ont développés une méthode TopGO en intégrant des connaissance à priori sur la structure du graphe GO. L'idée de cette approche est de crée un ensemble des annotations définit à partir de la liste des gènes d'intérêt, un parcours dans le graphe de GO sera effectué de bas en haut, et pour chaque terme on calcule

le p_value associé, si cette valeur est inférieure à la seuil fixé, alors on garde ce terme et on passe aux termes suivants, sinon on l'élimine de l'ensemble des termes construit au début, et par conséquent les gènes annotés par ce terme vont automatiquement le perdre. Par la suite, le nombre de gènes d'intérêt va changer en retirant les gènes qui n'avaient que les annotations avec forte p_value. Ce processus pourrait etre vue comme une étape de réduction de nombre des termes : Les termes avec une forte p_value sont éliminé du processus d'enrichissement, et par la suite va diminuer le biais généré à cause de leur présence dans l'ensemble des termes.
Cette méthode est implémentée dans le R package **topGO** (Alexa and Rahnenfuhrer,2016).
Une étude d'ontologies était effectuée pour les gènes différentiellement exprimés et les résultats obtenus sont résumés dans les tableaux suivant:

Ontology	e-val
regulation of biological quality	0.00013
regulation of post-embryonic development	0.00013
drug transmembrane transport	0.00016
systemic acquired resistance	0.00019
cellular response to oxygen-containing	0.00020
organelle organization	0.00020
flower development	0.00023
chlorophyll biosynthetic process	00.00024
rganic cyclic compound biosynthetic process	0.00025
peptidyl-L-cysteine S-palmitoylation	0.00026

Table 12.1: liste des 10 ontologies les plus exprimés pour les gènes différentiellement exprimés.

13 Annexe 3 :Réseau dynamique de réponse d'*Arabidopsis thaliana* au simulation sonore avec 8 noeuds:

13.1 Classification à $2-$signatures

La méthode de classification de variation de signe a été réalisée sur la liste des 9554 gènes différentiellement exprimés. Les gènes ont été regroupés sur la base de leurs variations temporelles en 8 groupes. L'allocation des gènes dans les 8 classes calculés a conduit à une distribution hétérogène de taille des classes qui varie entre 367 gènes et 2397 gènes (Figure 13.4.A). cette méthode de classification mène à une distribution gaussienne de chaque classe pour les quatre points temporels (0,1,3,8). Les distribution à chaque instant de la classe 1 sont présentés dans la figure 13.1. Cette distribution gaussienne était la même pour tous les autres classes.

Les signes associés à chaque groupe ainsi que le nombre des gènes sont présentés dans le tableau 13.1 :

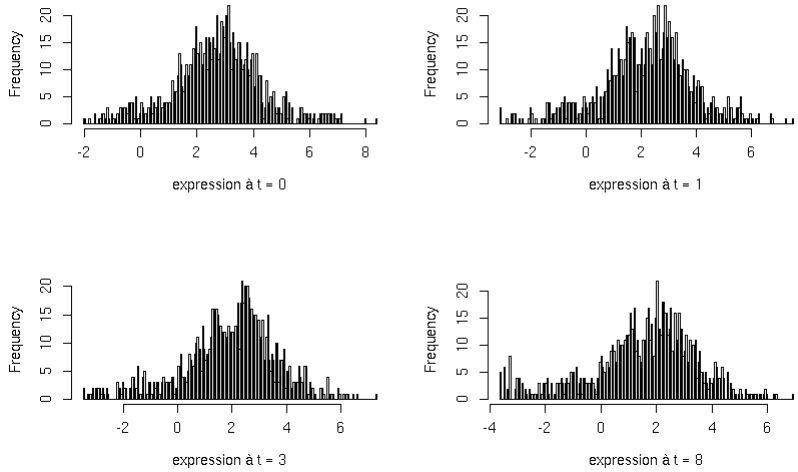


Figure 13.1: Histogramme des distributions à chaque instant de mesure pour la classe 1 obtenues en utilisant la méthode de classification avec deux signes.

classe	signe	nombre des gènes
1	- - -	1013
2	- - +	685
3	- + +	2397
4	+ + -	941
5	+ - -	1543
6	+ - +	1997
7	+ - -	610
8	+ + +	367

Table 13.1: Représentation des signes de chaque classe et le nombre des gènes associés.

Les 8 classes sont présentés dans les deux figures 13.2 et 13.3 en précision l'expression et la moyenne de chaque classe. Les 8 classes obtenus contiennent un grand nombre de gènes, ce qui conduit à une variabilité intra-classe importante. Pour modéliser les relations entre ces classes en inférant un réseau), la prise en compte de cette variabilité intra-classe sera une étape critique et c'est pour cela que nous avons utilisé le modèle (15.1) qui minimise l'erreur de prédiction pour le reste de l'analyse biologique.

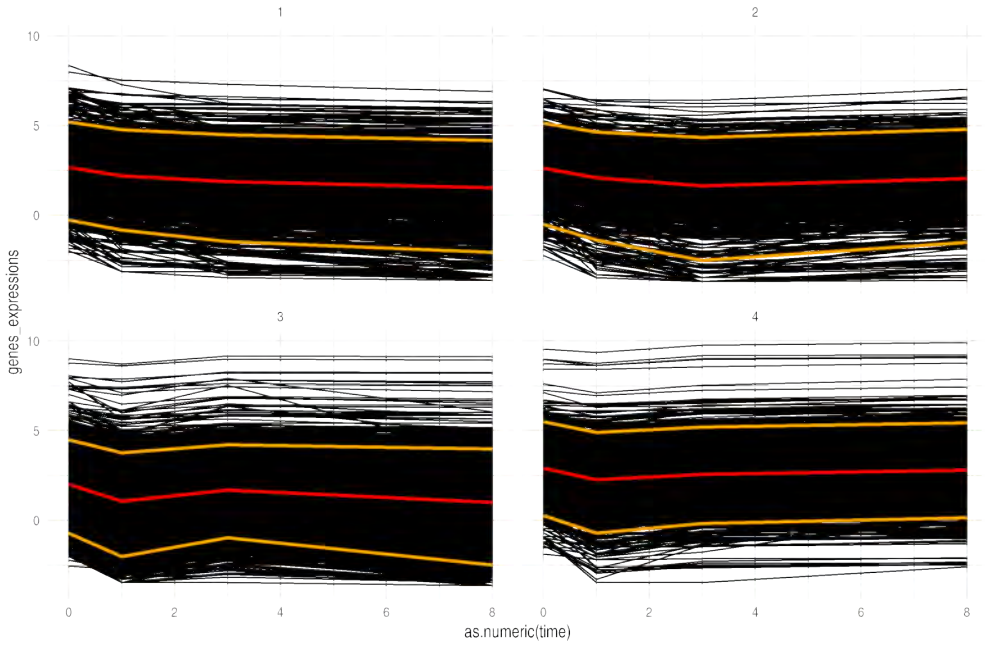


Figure 13.2: Représentation de résultat de classification des classes 1 à 4 obtenus en utilisant la méthode de classification des signes. la ligne rouge représente l'expression moyenne de chaque classe et les deux lignes en orange représentent les deux quantile à 95% et à 5%.

13.2 Réseau de régulation de l'effet de répétition sonore

Le réseau résultant décrit par la matrice A présente une légère structure hiérarchique composée de deux couches (Figure 13.4.B). La première couche est composée des classes 1 à 5 qui sont liées entre elles et se régulent et la deuxième couche est composée de la classe 6. La première couche est composée de classes hautement inter-connectés sans parents alors que la deuxième couche est composée d'un classe sans enfant. La plupart des liens sont répartis au sein de la couche 1 (87,5 %). Les boucles de régulation entre les classes ont été trouvées exclusivement dans la couche 1. L'analyse du vecteur b révèle que 4 classes sont plus sensibles au stimulations acoustiques (3,4,5,6). Le vecteur b est présenté dans la figure 13.5 et la figure 13.4. Cependant, chaque classe a est modulé par les stimulations acoustiques (Figure 13.4A, B).

Pour comprendre la signification biologique du réseau dynamique inféré, nous avons effectué une analyse d'enrichissement du processus biologique GO de chaque classe par rapport à tous les gènes composant le réseau. À l'exception du classe

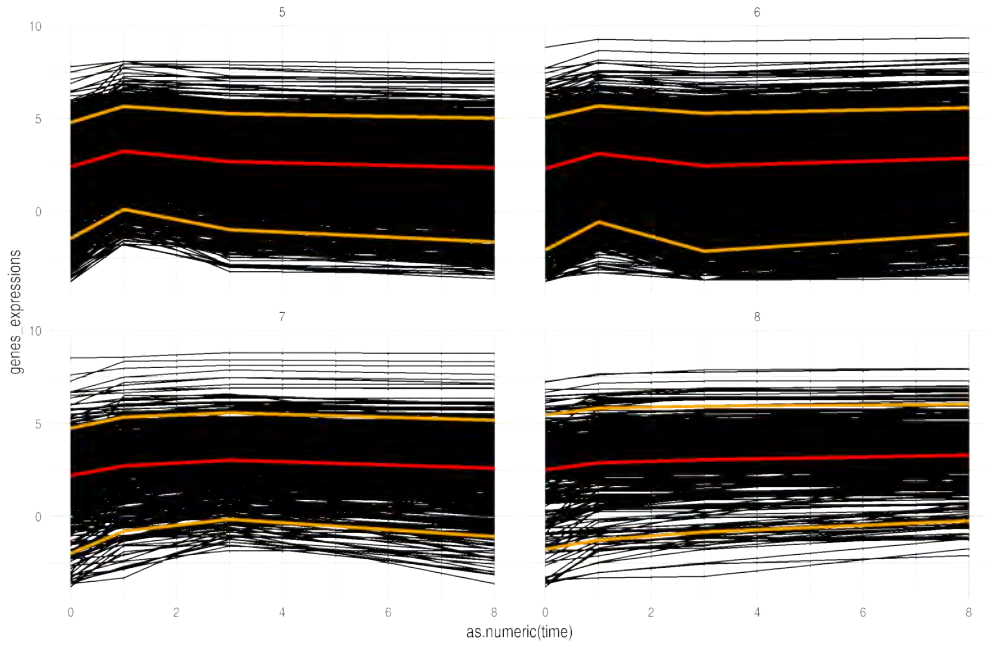


Figure 13.3: Représentation de résultat de classification des classes 5 à 8 obtenus en utilisant la méthode de classification des signes. la ligne rouge représente l'expression moyenne de chaque classe et les deux lignes en orange représentent les deux quantile à 95% et à 5%.

8, chaque groupe montre des enrichissements significatifs (liée à la résistance) (Figure 13.4.B). Ces résultats suggèrent que les stimulations acoustiques répétées affectent différentes fonctions inter-connectées de la plante conduisant à l'activation de l'immunité végétale.

13.3 Effet du son sur chaque classe

: Les stimulations acoustiques répétées affectent les expressions des gènes dans chaque groupe, régulant positivement les gènes des classes 1 à 4 et régulant négativement les gènes des classes 5 à 8 (Figure 13.5). 5937 (62 % de DEG) gènes des groupes 6, 5 et 3 étaient plus sensibles au son et activés dans Initiation de le priming du répétition sonore (13.4, B).Au total, ces résultats suggèrent que la répétition sonore affecte directement l'expression des 9554 gènes décrits par le réseau dynamique.

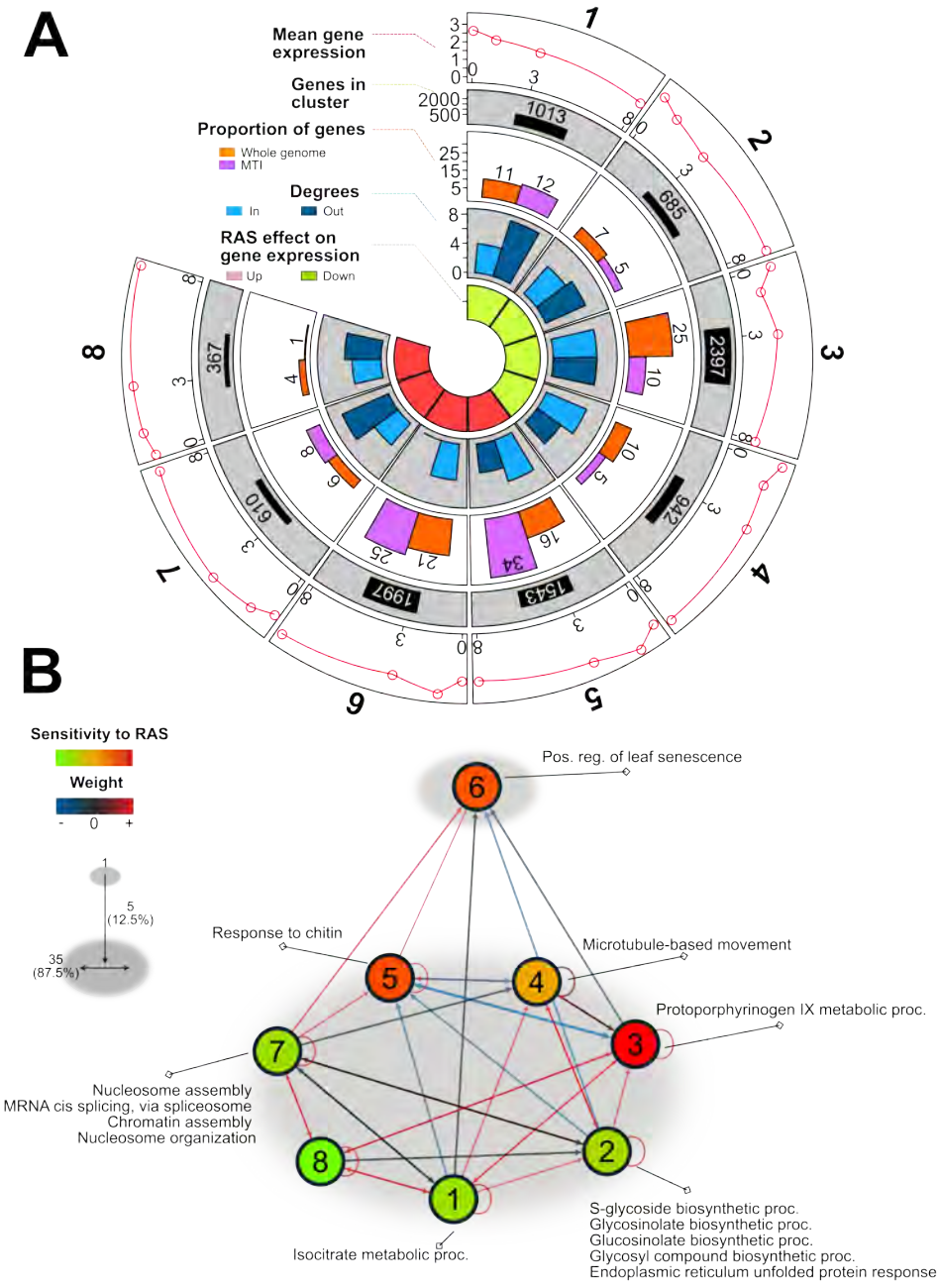


Figure 13.4: Réseau de gènes dynamique à l'échelle du génome. **A** Les gènes ont été classés par la méthode de classification à 2-signatures (moyenne de chaque classe dans la couche 1). Le nombre de gènes dans les classes variait d'un facteur 6 (couche 2) et représentait jusqu'à 25% du génome entier (couche 3). La structure de la matrice A indiquait un réseau structuré hiérarchiquement (couche 3) alors que b suggérait que les expressions de chaque classe était modulée par le son (couche 4). **B**. Le réseau dynamique inféré. Pour chaque classe , des processus biologiques GO significatifs sont indiqués.

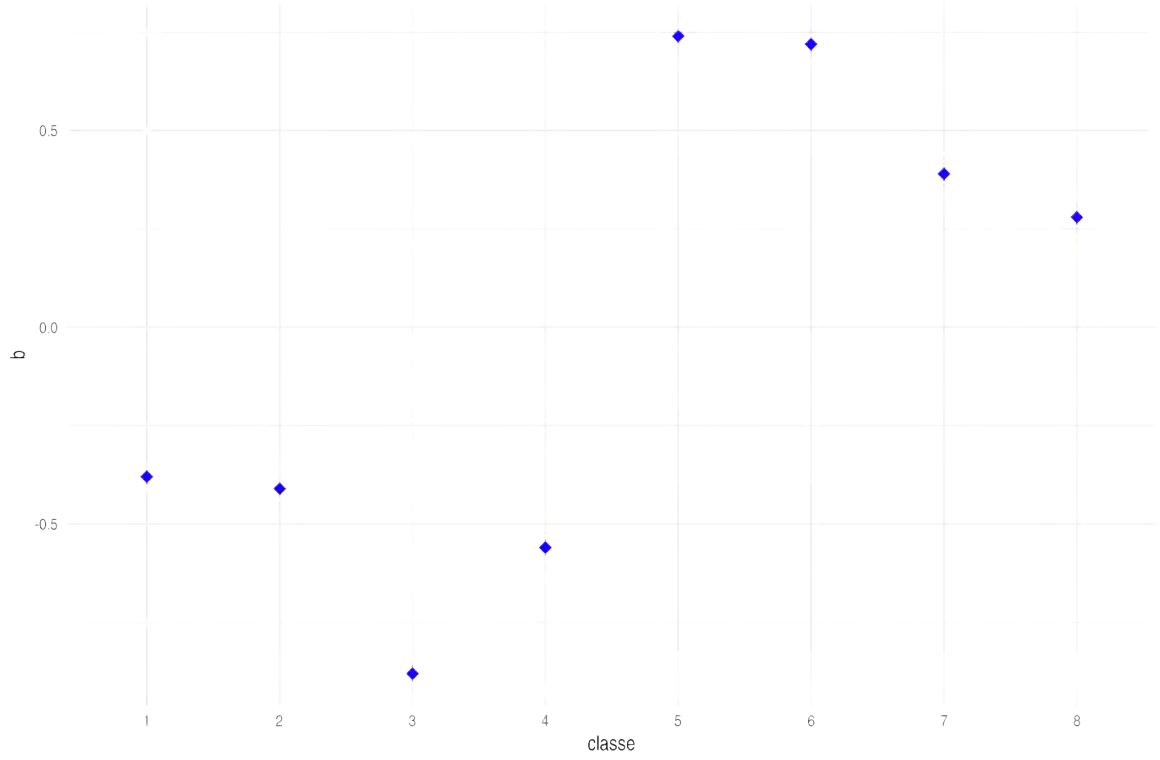


Figure 13.5: L'expression de b prédite par le modèle utilisé, qui décrit l'effet du simulation sonore sur chaque classe.

13.4 Stabilité de réseau

Pour étudier la stabilité du réseau, on étudie les valeurs propres de la matrice des coefficients A. Pour calculer les modes qui vont nous permettre de trouver en combien du temps le système retourne à l'équilibre en l'absence de stimulation, on commence par calculer les valeurs propres et puis on étudie les signes des parties réelles associées.

L'ensemble des valeurs propres est composé de cinqs valeurs propres réelles (-0.013 , 0.000, 0.259, 0.259 et -0.834) et complexes (-0.153+0.031i , -0.153-0.031i et 0.053+0.034i) .

Si les parties réelles sont négatives ou nulles alors elles correspondent à des modes stables. Pour trouver le temps de retour à l'équilibre il suffit de calculer l'inverse de la partie réelle des valeurs propres à partie réelle négative. Après l'analyse des valeurs propres nous avons constaté que 5 des 8 modes sont des modes stables. Le mode principal atteint l'état initial 29h (1.12 jours) après le dernier exposition sonore, tandis que les modes 2 et 3 atteignent l'équilibre après 156h (6,5 jours). Cette analyse des valeurs propres montre qu'une plante primé exposée à 3 stimulations acoustiques oublierait le priming après 1.12 jours.

14 Annexe 4 : Intégration des données phénotypiques

Sommaire

Parmi les mesures obtenues au début de thèse, nous avons des mesures phénotypiques (croissance de champignon) ainsi que des mesures transcriptomiques de la plante. Par l'expression Intégration de phénotype nous voulons chercher et explorer les changements d'expression de gènes (moyennes des classes) associés avec un changement des valeurs de phénotype de champignon *S. sclerotiorum*. Cette étape d'intégration sera la dernière étape avant l'obtention de réseau final où on trouve l'expression de phénotype liée au réseau d'expression des gènes. La figure 14.1 décrit les différentes étapes utilisées pour avoir ce réseau final.

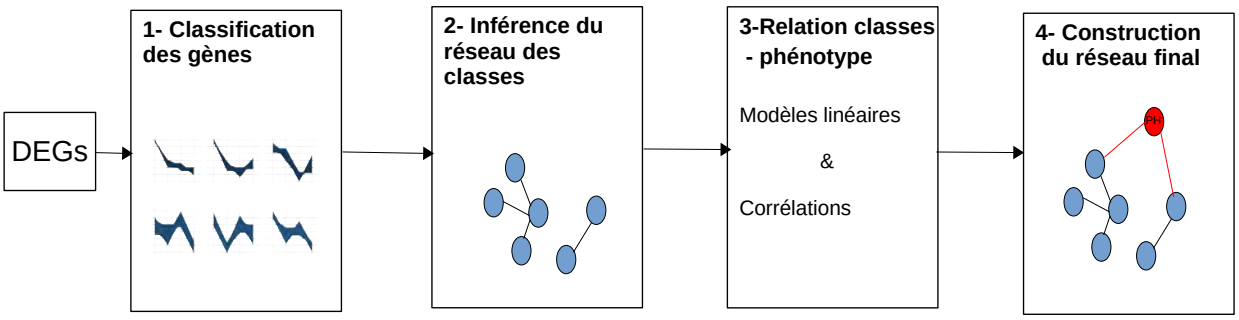


Figure 14.1: Les différentes étapes de construction de réseau reliant l'expression de phénotype aux expressions de gènes.

Dans ce chapitre nous essayons d'établir un lien direct entre les mesures d'expressions des gènes et l'expression de phénotype. Deux méthodes ont été utilisées pour atteindre ce but ; un modèle linéaire généralisé sur les expressions moyenne de chaque classe et des corrélations simples. Les différentes étapes utilisées sont présentées dans ce chapitre.

14.1 Méthodes d'intégration

Pour établir le lien entre les expressions phénotypiques et nos classes, nous pouvons utiliser deux méthodes différente s:

Utilisation d'un modèle linéaire:

Les relations entre les gènes et le phénotype PH ont été estimées à l'aide d'un modèle linaire classique. En utilisant ce modèle nous n'avons pas utilisé le temps explicitement comme paramètre dans le modèle et c'est parce que nous n'avons pas des séries temporelles longues (4 points pour le phénotype aussi). Le modèle s'écrit sous la forme:

$$PH = \mu + \beta_1 C_1 + ... + \beta_k C_k + \epsilon \tag{14.1}$$

où PH est le phénotype qu'on veut l'exprimer et C_k, est l'expression moyenne de chaque classe avec k est le nombre des classes. et ϵ est le terme d'erreur du modèle qui suit une loi normale.
les relations entre phénotype et classe sont évalués en effectuant le test :

$$H_0 : \beta_k = 0 \qquad contre \qquad H_1 : \beta_k \neq 0 \tag{14.2}$$

Des corrections de tests multiples (Benjamini and Hochberg,1995) sont ensuite été réalisées. Le lien entre une classe et le phénotype est considéré comme significative si sa p-valeur ajustée est inférieure à 5%.
Si nous avons un grand nombre des classe, avant de commencer cette étape d'intégration, nous pouvons appliquer une ACP sur les expressions moyennes des classes, pour choisir les classes les plus exprimées selon le premier axe, et puis d'établir une relation entre ces classes et l'expression de phénotype en utilisant le modèle linéaire définit précédemment.

14.1.1 Utilisation des corrélations:

La plus simple méthode pour trouver un lien entre un ensemble des gènes (classe des gènes) et le phénotype est d'utiliser les corrélations. Il suffit de calculer les corrélations (de spearman) entre les 8 classes et le phénotype puis fixer un seuil qu'à partir de laquelle on peut considérer que la relation entre le gène et le phénotype est significative. Le lien entre une classe et le phénotype est considéré comme significative si le coefficient de corrélation est supérieur en valeur absolue à le seuil fixé.

Intégration par un modèle linéaire

Nous avons appliqué le modèle linéaire directement sur les expressions moyennes de chaque classe.

Le problème c'est que le modèle ne peut estimer que seulement trois coefficients parmi le 8 coefficients de chaque classe (Sortie R 14.1), donc nous avons décidé qu'avant d'appliquer le modèle linéaire d'appliquer une ACP sur les expressions moyenne de chaque classe. Le but de cette étape est de réduire le nombre de classe de 8 à 3 classes (vu que le modèle ne peut estimer que trois coefficients) pour avoir les classes qui sont le plus exprimé par rapport aux autres classe et qui vont être utilisées comme des entrées dans le modèle linéaire.

Listing 14.1: résultat du modèle linéaire généralisé

```
  Call  :
glm(formula = ph_int ~ c1 + c2 + c3 + c4 + c5 + c6 + c7 + c8, family = gaussian)

Deviance Residuals:
          1           2           3           4           5           6
          7           8
 0.0000000  -0.0005058   0.0010117  -0.0003035  -0.0002023  -0.0001012
   0.0000000   0.0001012

Coefficients: (5 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0125332  0.0015810   7.928  0.00137 **
c1           0.0099637  0.0012945   7.697  0.00153 **
c2          -0.0006586  0.0012847  -0.513  0.63522
c3          -0.0039642  0.0009367  -4.232  0.01335 *
c4                  NA         NA      NA       NA
c5                  NA         NA      NA       NA
c6                  NA         NA      NA       NA
c7                  NA         NA      NA       NA
c8                  NA         NA      NA       NA
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1
  1
```

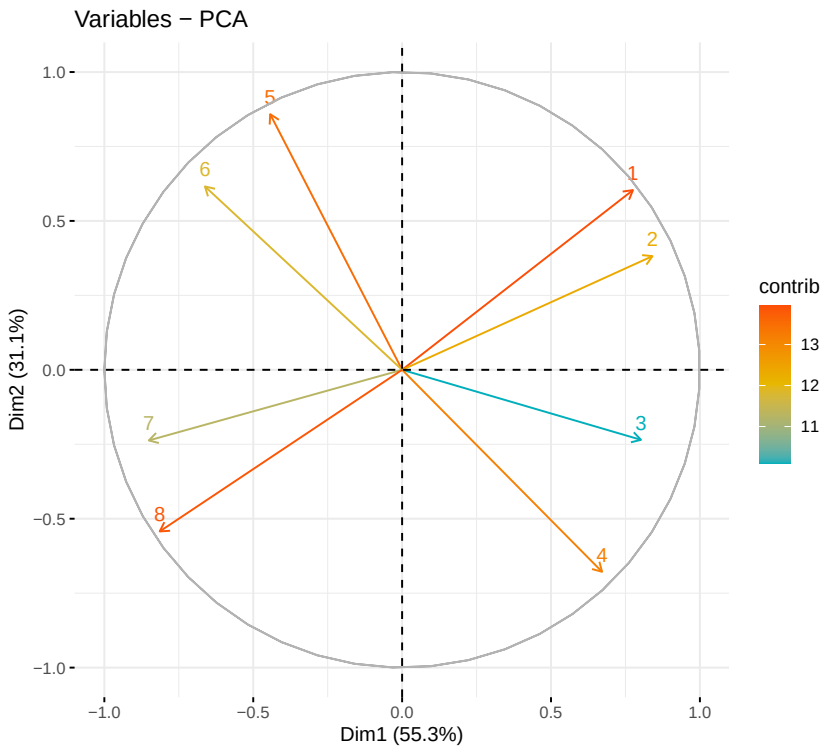


Figure 14.2: Projection des classes (de 1 à 8) sur les deux premiers axes de l'ACP.

Les résultats de cette étape nous a permis de travailler avec les trois classes 7, 2 et 8 (Figure 14.3). Pour être sur de ce choix nous avons calculé l'erreur de

prédiction de phénotype en utilisant différentes combinaisons des classes. L'erreur de prédiction était le même pour toutes les combinaisons testé (Sortie R 14.2).

Listing 14.2: Erreur de prédiction en utilisant des différentes combinaisons des classes

```
1  "";"x.c1";"x.c2";"x.c3";"x.erreur"
2  "1";3;6;8;1,791118117993e-07
3  "2";1;8;6;1,79111811799299e-07
4  "3";4;7;8;1,79111811799299e-07
5  "4";8;6;1;1,79111811799299e-07
6  "5";6;1;2;1,79111811799301e-07
7  "6";1;7;3;1,791118117993e-07
8  "7";5;7;8;1,791118117993e-07
9  "8";3;5;1;1,791118117993e-07
10 "9";7;2;5;1,79111811799302e-07
11 "10";2;7;8;1,79111811799302e-07
```

En effet, ces résultats ne sont pas choquants par le fait que le modèle nous a permis de déduire que les moyennes des classes sont fortement corrélées les unes aux autres, de sorte qu'elles ne fournissent pas d'informations uniques ou indépendantes dans le modèle de régression (ce que explique le coefficient NA trouvés). Pour résoudre ce problème, nous pouvons simplement supprimer l'une des variables corrélées du modèle car elles ne fournissent pas réellement d'informations uniques ou indépendantes dans le modèle de régression. Le modèle final contiendra la même estimation de coefficient pour les variables que nous décidons de conserver et la qualité globale de l'ajustement du modèle sera la même (ce que explique le même erreur du modèle en utilisant des différentes combinaisons). On peut expliquer aussi, ce problème de modélisation par le fait que le système engendré par l'ensemble des moyennes de classes est de rang trois, et donc quelle que soit la combinaison que nous allons le choisir on va trouver toujours que le phénotype est exprimé seulement avec 3 classes. En conclusion, ce type de modèle n'est pas le meilleur pour décrire la relation entre le phénotype et les classes, et on peut expliquer ça par le fait d'avoir seulement quatre points de mesures.

14.2 Intégration par des corrélations

Nous avons calculé la corrélation de Spearman entre l'expression moyenne et l'expression de phénotype (Figure 14.3). Cette méthode n'est pas trop conseillée puisqu'elle impose l'existence des corrélations indirecte. Les corrélations entre le phénotype et les différents moyennes des classes variées entre -0.8978274 et 0.9256324 (Figure 14.3). En prenant en compte les corrélations obtenues en valeur absolue, nous pouvons déduire que le phénotype est corrélé le plus avec la classe 1

et la classe 8 en fixant le seuil à 0.8.

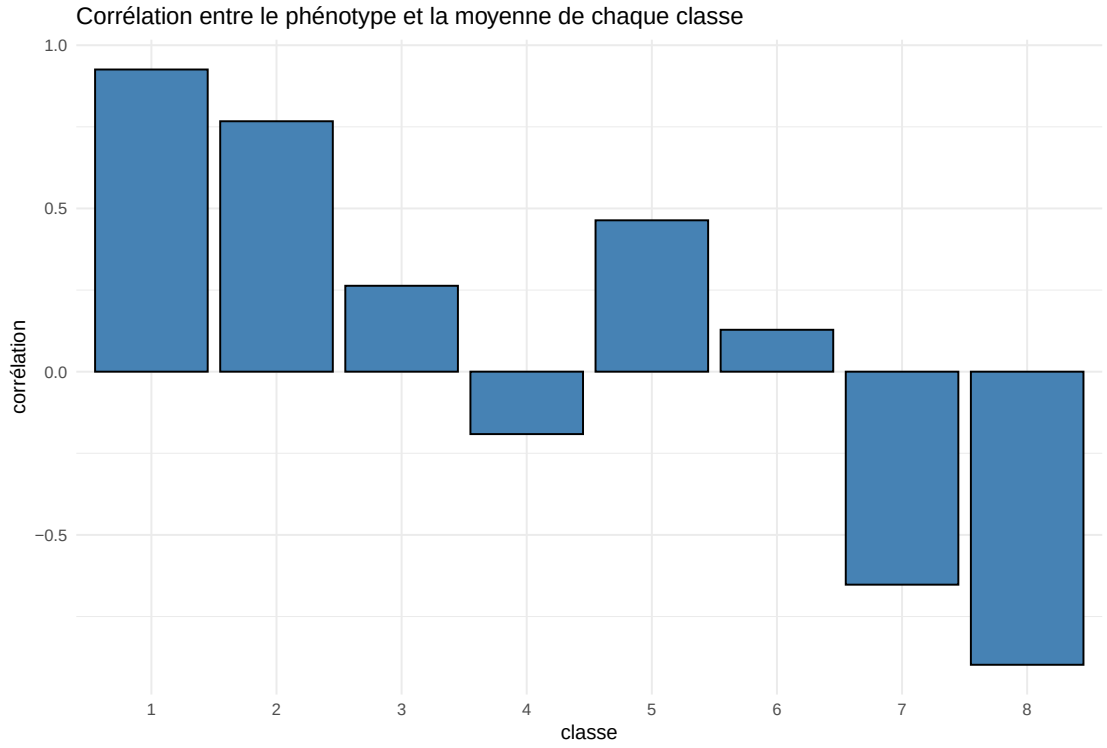


Figure 14.3: corrélation entre la valeur moyenne de chaque classe et le phénotype.

En conclusion, l'utilisation d'un modèle linéaire généralisé et des corrélations classiques sur les expressions moyenne de chaque classe peut nous permettre d'inférer un réseau final en intégrant l'expression phénotypique dedans. Mais ces deux modèles n'arrivent pas vraiment à choisir les 'meilleurs' contributeurs à l'expression phénotypique. Nous pouvons expliquer ce problème de modélisation par le fait qu'on a pas assez des points de mesures pour l'expression de gènes ni pour l'expression phénotypique.

15 Annexe 5 : Analyse et illustration des formules utilisées

Propriété

Le modèle que nous avons retenu pour la modélisation est défini par l'équation suivante :

$$\begin{aligned} X_{g,t+1} - X_{g,t} &= A_{i,i}X_{g,t} + \sum_{j\neq i} A_{i,j}X_{j,t}^{\alpha_{g,t}} + b_i + \epsilon_t, \quad \text{pour tout } t = 0, \cdots T-1, \\ X_{g,1} - X_{g,0} &= b_i + \epsilon_0, \quad \text{pour tout } t = 0, \cdots T-1, \quad i = 1, \cdots K, \quad g \in C_i, \end{aligned} \tag{15.1}$$

où $\alpha_{g,t} = Pr[X_{g',t} \leq X_{g,t},\ g' \in C_j]$ représente l'ordre du quantile $X_{g,t}$.

Si nous avons ce modèle, alors nous savons qu'il est également applicable avec les expressions moyennes.

Preuve

Pour dériver une équation représentant l'évolution des moyennes de chaque classe ($\bar{X}_t$), nous pouvons utiliser le fait que la moyenne d'un ensemble de données est simplement la somme des valeurs divisée par le nombre de valeurs.

Si nous avons N valeurs $X_{g,t}$ dans une classe i à la période t, alors la moyenne de cette classe à cette période est :

$$\bar{X}_{i,t} = \frac{1}{N} \sum_{g \in C_i} X_{g,t}$$

Maintenant, si nous dérivons l'évolution de $\bar{X}_{i,t}$ d'une période à l'autre ($\bar{X}_{i,t+1} - \bar{X}_{i,t}$), nous pouvons utiliser la même logique que celle utilisée pour les valeurs individuelles $X_{g,t}$ en remplaçant les valeurs par les moyennes.

En utilisant la notation $\bar{X}_{i,t}$ pour représenter la moyenne de la classe i à la période t, nous pouvons réécrire l'équation donnée comme suit :

$$\bar{X}_{i,t+1} - \bar{X}_{i,t} = \frac{1}{N} \sum_{g \in C_i} (X_{g,t+1} - X_{g,t})$$

Maintenant, en utilisant l'équation donnée :

$$X_{g,t+1} - X_{g,t} = A_{i,i} X_{g,t} + \sum_{j \neq i} A_{i,j} X_{j,t}^{\alpha_{g,t}} + b_i + \epsilon_t$$

Nous pouvons substituer cette expression dans l'équation pour $\bar{X}_{i,t+1} - \bar{X}_{i,t}$:

$$\begin{aligned}
\bar{X}_{i,t+1} - \bar{X}_{i,t} &= \frac{1}{N} \sum_{g \in C_i} (A_{i,i} X_{g,t} + \sum_{j \neq i} A_{i,j} X_{j,t}^{\alpha_{g,t}} + b_i + \epsilon_t) \\
&= A_{i,i} \bar{X}_{i,t} + \frac{1}{N} \sum_{g \in C_i} \sum_{j \neq i} A_{i,j} X_{j,t}^{\alpha_{g,t}} + b_i + \frac{1}{N} \sum_{g \in C_i} \epsilon_t \\
&= A_{i,i} \bar{X}_{i,t} + \sum_{j \neq i} A_{i,j} \left(\frac{1}{N} \sum_{g \in C_i} X_{j,t}^{\alpha_{g,t}} \right) + b_i + \frac{1}{N} \sum_{g \in C_i} \epsilon_t
\end{aligned}$$

Alors on a :

$$\begin{aligned}
\bar{X}_{i,t+1} - \bar{X}_{i,t} &= A_{i,i} \bar{X}_{i,t} + \sum_{j \neq i} A_{i,j} \left(\frac{1}{N} \sum_{g \in C_i} X_{j,t}^{\alpha_{g,t}} \right) + b_i + \frac{1}{N} \sum_{g \in C_i} \epsilon_t \\
&= A_{i,i} \bar{X}_{i,t} + \sum_{j \neq i} \left(A_{i,j} \frac{1}{N} \sum_{g \in C_i} X_{j,t}^{\alpha_{g,t}} \right) + b_i + \frac{1}{N} \sum_{g \in C_i} \epsilon_t \\
&= A_{i,i} \bar{X}_{i,t} + \sum_{j \neq i} \left(\frac{1}{N} \sum_{g \in C_i} A_{i,j} X_{j,t}^{\alpha_{g,t}} \right) + b_i + \frac{1}{N} \sum_{g \in C_i} \epsilon_t \\
&= A_{i,i} \bar{X}_{i,t} + \sum_{j \neq i} \left(\frac{1}{N} \sum_{g \in C_i} A_{i,j} X_{j,t}^{\alpha_{g,t}} \right) + b_i + \frac{1}{N} \sum_{g \in C_i} \epsilon_t
\end{aligned}$$

Propriété

Si les distributions d'expressions de gènes au sein des classes suivent une lois normale $X_{g,t} \sim \mathcal{N}(\bar{X}_t^k, \sigma_t^k)$, le modèle 15.1 devient:

$$X_{g,t+1} - X_{g,t} = \theta_{i,t}(X_{g,t} - \bar{X}_t^i) + \sum_j A_{i,j}\bar{X}_t^j + b_i + \epsilon_t, \tag{15.2}$$

avec $\theta_{i,t} = \sum_j \frac{\sigma_t^j}{\sigma_t^i} A_{i,j}$.

Preuve:

supposons que les expressions géniques au sein des classes suivent une loi normale :

$$X_{g,t} \sim \mathcal{N}(\bar{X}_t^k, \sigma_t^k)$$

Cela signifie que nous pouvons exprimer $X_{g,t}$ comme un échantillon de la distribution normale pour la classe à ce moment donné :

$$X_{g,t} = \bar{X}_t^i + \sigma_t^i \epsilon_{g,t}$$

où $\epsilon_{g,t}$ est un échantillon d'une distribution normale centrée réduite.

En substituant cette expression dans le modèle précédent (équation 15.1), nous obtenons :

$$X_{g,t+1} - X_{g,t} = \theta_{i,t}\sigma_t^i \epsilon_{g,t} + \sum_j A_{i,j}\bar{X}_t^j + b_i + \epsilon_t$$

Ce qui conduit au modèle suivant :

$$X_{g,t+1} - X_{g,t} = \theta_{i,t}(X_{g,t} - \bar{X}_t^i) + \sum_j A_{i,j}\bar{X}_t^j + b_i + \epsilon_t \tag{15.3}$$

avec $\theta_{i,t} = \sum_j \frac{\sigma_t^j}{\sigma_t^i} A_{i,j}$.

Propriété : Erreur de prédiction

l'erreur quadratique moyenne normalisée de prédiction (NMSE) est défini comme :

$$err = \sum_{m=2}^{M} \|X_{t_m} - \tilde{A}\hat{X}_{t_m-1}\|^2 / \sum_{m=2}^{M} \|X_{t_m}\|^2 \tag{15.4}$$

où X_{t_m} représente le vecteur d'expression des N gènes observé à l'instant t_m, $\hat{X}_{t_m-1}$ le vecteur des expressions moyennes prédites ou observées à l'instant précédent $t_m - 1$, et où $\tilde{A}$ est définie comme une matrice par bloc de dimension $N \times N$ telle que $\tilde{a}_{i,j} = \frac{a_{k,l}}{N_l}$ pour un gène i dans le cluster C_k et le gène j dans le cluster C_l. L'idée ici est d'intégrer dans l'erreur de prédiction l'effet du clustering. Puisque la structure de $\tilde{A}$ est spéciale (par bloc), nous pouvons reformuler l'expression de l'erreur comme :

$$err = 1 - \frac{2Tr(CAV) - Tr(CASA')}{Tr(W)} \tag{15.5}$$

où C est la matrice diagonale $C_{kk} = N_k$, $k = 1, K$, et V, W et S sont les matrice de taille $K \times K$ de la variance-covariance empirique $V = \bar{X}_{-M}\bar{X}'_{-1}$, $W = \bar{X}_{-1}\bar{X}'_{-1}$ et $S = \bar{X}_{-M}\bar{X}'_{-M}$, où $\bar{X}_{-l}$ désigne la matrice de données moyennées $\bar{X}$ privée de sa colonne l.

On veut montrer l'equivalence entre les deux formules de l'erreur.

Preuve

Nous commençons avec l'expression de l'erreur normalisée de prédiction (err) :

$$\text{err} = \frac{\sum_{m=2}^{M} \|X_{t_m} - \tilde{A}\hat{X}_{t_m-1}\|^2}{\sum_{m=2}^{M} \|X_{t_m}\|^2}$$

Nous voulons réécrire cette expression en utilisant les moyennes des données ($\bar{X}_{-1}$) et la covariance (W). Pour cela, nous allons développer le numérateur et le dénominateur de cette expression.

Développons le numérateur :

$$\sum_{m=2}^{M} \|X_{t_m} - \tilde{A}\hat{X}_{t_m-1}\|^2$$

En utilisant les moyennes ($\bar{X}_{-1}$) et la covariance (W), nous pouvons réécrire $X_{t_m} - \tilde{A}\hat{X}_{t_m-1}$ comme $(X_{t_m} - \bar{X}_{-1})(\hat{X}_{t_m-1} - \bar{X}_{-1})'$. Ainsi, le numérateur devient :

$$\begin{aligned}\sum_{m=2}^{M} \|X_{t_m} - \tilde{A}\hat{X}_{t_m-1}\|^2 &= \sum_{m=2}^{M} \|(X_{t_m} - \bar{X}_{-1})(\hat{X}_{t_m-1} - \bar{X}_{-1})'\|^2 \\ &= \sum_{m=2}^{M} (X_{t_m} - \bar{X}_{-1})(\hat{X}_{t_m-1} - \bar{X}_{-1})'(X_{t_m} - \bar{X}_{-1})(\hat{X}_{t_m-1} - \bar{X}_{-1})'\end{aligned}$$

Maintenant, développons le dénominateur :

$$\sum_{m=2}^{M} \|X_{t_m}\|^2$$

En utilisant également les moyennes ($\bar{X}_{-1}$), le dénominateur peut être réécrit comme $\sum_{m=2}^{M} \|(X_{t_m} - \bar{X}_{-1})\|^2$.

Revenons à notre expression d'erreur (err) :

$$\text{err} = \frac{\sum_{m=2}^{M}(X_{t_m} - \bar{X}_{-1})(\hat{X}_{t_m-1} - \bar{X}_{-1})'(X_{t_m} - \bar{X}_{-1})(\hat{X}_{t_m-1} - \bar{X}_{-1})'}{\sum_{m=2}^{M} \|(X_{t_m} - \bar{X}_{-1})\|^2}$$

En divisant le numérateur et le dénominateur par $\text{Tr}(W)$, nous pouvons simplifier davantage cette expression.

Continuons la simplification en divisant le numérateur et le dénominateur par $\text{Tr}(W)$, la trace de la matrice de covariance empirique :

$$\text{err} = \frac{\frac{1}{\text{Tr}(W)}\sum_{m=2}^{M}(X_{t_m} - \bar{X}_{-1})(\hat{X}_{t_m-1} - \bar{X}_{-1})'(X_{t_m} - \bar{X}_{-1})(\hat{X}_{t_m-1} - \bar{X}_{-1})'}{\frac{1}{\text{Tr}(W)}\sum_{m=2}^{M} \|(X_{t_m} - \bar{X}_{-1})\|^2}$$

Cela nous permet d'exprimer l'erreur en termes de la variance-covariance empirique (W).

En effectuant cette division, nous avons maintenant :

$$\text{err} = \frac{\sum_{m=2}^{M}(X_{t_m} - \bar{X}_{-1})(\hat{X}_{t_m-1} - \bar{X}_{-1})'(X_{t_m} - \bar{X}_{-1})(\hat{X}_{t_m-1} - \bar{X}_{-1})'}{\sum_{m=2}^{M} \|(X_{t_m} - \bar{X}_{-1})\|^2}$$

Pour simplifier davantage l'expression, réarrangeons les termes pour identifier les contributions de CAV, $CASA'$, et W.

$$\text{err} = \frac{\sum_{m=2}^{M}(X_{t_m} - \bar{X}_{-1})(\hat{X}_{t_m-1} - \bar{X}_{-1})'(X_{t_m} - \bar{X}_{-1})(\hat{X}_{t_m-1} - \bar{X}_{-1})'}{\sum_{m=2}^{M} \|(X_{t_m} - \bar{X}_{-1})\|^2}$$

Nous pouvons réécrire $X_{t_m} - \bar{X}_{-1}$ comme $X_{t_m} - \bar{X}_{-1} = (X_{t_m} - \bar{X}_{-1}) - \bar{X}_{-1}$, où $\bar{X}_{-1}$ est la moyenne des données. Ensuite, nous obtenons :

$$\begin{aligned}\text{err} &= \frac{\sum_{m=2}^{M}((X_{t_m} - \bar{X}_{-1}) - \bar{X}_{-1})(\hat{X}_{t_m-1} - \bar{X}_{-1})'((X_{t_m} - \bar{X}_{-1}) - \bar{X}_{-1})(\hat{X}_{t_m-1} - \bar{X}_{-1})'}{\sum_{m=2}^{M} \|(X_{t_m} - \bar{X}_{-1})\|^2} \\ &= \frac{\sum_{m=2}^{M}(X_{t_m} - \bar{X}_{-1})(\hat{X}_{t_m-1} - \bar{X}_{-1})' - (X_{t_m} - \bar{X}_{-1})\bar{X}'_{-1} - \bar{X}_{-1}(\hat{X}_{t_m-1} - \bar{X}_{-1})' + \bar{X}_{-1}\bar{X}'_{-1})}{\sum_{m=2}^{M} \|(X_{t_m} - \bar{X}_{-1})\|^2}\end{aligned}$$

Nous avons maintenant exprimé l'erreur en termes de moyennes et de covariances des données. Nous pouvons identifier les contributions de chaque terme à CAV,

$CASA'$, et W.

Maintenant, réorganisons les termes pour identifier les contributions de CAV, $CASA'$, et W :

$$\text{err} = \frac{\sum_{m=2}^{M}(X_{t_m} - \bar{X}_{-1})(\hat{X}_{t_m-1} - \bar{X}_{-1})' - (X_{t_m} - \bar{X}_{-1})\bar{X}'_{-1} - \bar{X}_{-1}(\hat{X}_{t_m-1} - \bar{X}_{-1})' + \bar{X}_{-1}\bar{X}'_{-1}}{\sum_{m=2}^{M} \|(X_{t_m} - \bar{X}_{-1})\|^2}$$
$$= \frac{\sum_{m=2}^{M}(X_{t_m} - \bar{X}_{-1})(\hat{X}_{t_m-1} - \bar{X}_{-1})' - (X_{t_m}\bar{X}'_{-1} - \bar{X}_{-1}\bar{X}'_{-1} - \bar{X}_{-1}(\hat{X}_{t_m-1} - \bar{X}_{-1})' + \bar{X}_{-1}\bar{X}'_{-1}}{\sum_{m=2}^{M} \|(X_{t_m} - \bar{X}_{-1})\|^2}$$

Maintenant, identifions chaque terme :

- $(X_{t_m} - \bar{X}_{-1})(\hat{X}_{t_m-1} - \bar{X}_{-1})'$ représente la contribution de CAV, la covariance des prédictions par cluster. - $X_{t_m}\bar{X}'_{-1} - \bar{X}_{-1}\bar{X}'_{-1}$ représente la contribution de $CASA'$, la covariance des résidus par cluster. - Enfin, nous avons identifié $\bar{X}_{-1}\bar{X}'_{-1}$ comme contribuant à W, la variance totale des données.

Ainsi, nous avons réorganisé les termes pour isoler les contributions de CAV, $CASA'$, et W. Continuons dans la prochaine réponse pour résumer cette expression de manière plus compacte.

Maintenant, résumons l'expression en identifiant les contributions de CAV, $CASA'$, et W de manière plus compacte :

$$\text{err} = \frac{\sum_{m=2}^{M} \text{Termes de } CAV - \text{Termes de } CASA' + \text{Termes de } W}{\sum_{m=2}^{M} \text{Termes de } W}$$

En regroupant les termes, nous obtenons :

$$\text{err} = 1 - 2\frac{\text{Tr}(CAV) - \text{Tr}(CASA')}{\text{Tr}(W)}$$

Dans cette expression, $\text{Tr}(CAV)$ représente la trace de la covariance des prédictions par cluster, $\text{Tr}(CASA')$ représente la trace de la covariance des résidus par cluster, et $\text{Tr}(W)$ représente la trace de la variance totale des données.

Pour conclure, nous avons réarrangé l'expression de l'erreur pour exprimer l'erreur en fonction de CAV, $CASA'$, et W, ce qui nous permet de mieux comprendre les contributions de chaque composante à l'erreur de prédiction.

16 Productions scientifiques

Au cours de la thèse, en plus de l'article en cours de finalisation et de soumission, les différents résultats obtenus ont été présentés sous forme de deux posters et d'un flash talk, lors de deux conférences distinctes :

JOBIM 2022 : Journée Ouvertes en Biologie, Informatique et Mathématiques - Rennes
Dynamic genes network inference and very short time series. The case of the priming of plant immunity by repeated acoustic stimuli.
Khaoula Hadj-Amor, Adelin Barbacci, Frédérick Garcia

PBM 2022 : 10th Plant Biomechanics Conference - Lyon
How repeated acoustic stimuli increase Arabidopsis resistance to the necrotrophic fungus Sclerotinia sclerotiorum?
Khaoula Hadj Amor , Frédérick Garcia , Ophélie Léger , Mehdi Khafif , Sylvain Raffaele , Adelin Barbacci

Dynamic genes network inference and very short time series. How repeated acoustic stimuli affect plant immunity ?

Khaoula Hadj-Amor[1], Adelin Barbacci[2] and Frédérick Garcia[1]

[1] unité de Mathématiques et Informatique Appliquées de Toulouse (MIAT) , Institut National de Recherche pour l'Agriculture, l'alimentation et l'Environnement (INRAe), 24 Chemin de Borde-Rouge, 31326, Castanet-Tolosan, FRANCE

[2] Laboratoire des Interactions Plantes Microorganismes et Environnement (LIPME) , Institut National de Recherche pour l'Agriculture, l'alimentation et l'Environnement (INRAe) - Centre National de la Recherche Scientifique (CNRS), 24 Chemin de Borde Rouge, 31326, Castanet-Tolosan, FRANCE

Corresponding author: khaoula.hadj-amor@inrae.fr

1 Introduction

RNA-sequencing methods are central to biology and the basis of many breakthroughs in the understanding of the transcriptome. Also, their relatively moderate cost now enables the acquisition of transcriptome time series, which are useful to understand the dynamics of a biological phenomenon after a stimulus. However, The high dimensionality of RNA-seq data makes it hard to understand the complex dynamics that exists between measured genes.

Dynamic network inference methods are methods of choice to integrate the complexity of time-dependant transcriptomic data [1]: these methods produce networks in which nodes are genes and an edge between two genes represents a action (activation or repression) of one gene on the other during the time series. However, most network inference methods are also limited by the large dimension of RNA-seq data, in which the number of observations is always much lower than the number of measured gene expressions [2]. Hence, data dimension is frequently reduced by performing the inference on a small number of genes, picked *a priori*. In this presentation, we propose another strategy, which consists in grouping genes with similar expression dynamics to form a clustered gene network. More precisely, our clustered gene network inference strategy is a two step method (clustering then network inference). The clustering step is based on the analysis of sign variations in order to make it suitable for very-short time series.

Our strategy is illustrated on an original transcriptomic data associated with the priming of plant immune resistance by repeated acoustic stimuli (RAS).

2 Results

Arabidopsis Thaliana plants were exposed to sound three hours per day. For every sample, the number of reads corresponding to 28775 genes was reported. 0, 1, 8 days of RAS were described by quadruplicate samples whereas 3 RAS was described in triplicate. [PAS claire cette phrase]

A prior differential analysis showed that RAS modulate 7,108 genes involved in the resistance against the necrotrophic fungus *Sclerotinia sclerotiorum*. We restricted our analysis to these genes. The resulting clustering exhibited the lowest distance within clusters and the lowest Davies Bouldin index with a reduced computational time compared to other classical methods such as; Kmeans, kml, hierarchical classification, ... This clustering allows the inference of a reliable clustered gene network using a first order autoregressive model [3].

The resulting network has an hierarchical structure. In particular, a mechanosensitive channel, which is involved in mechano-perception, was shown to be associated with one of the main nodes of the network and to directly regulate others nodes (clusters of genes).

References

[1] Christopher A Penfold and David L Wild. How to infer gene networks from expression profiles, revisited. *Interface focus*, 1(6):857–870, 2011.

[2] Zhide Fang, Jeffrey Martin, and Zhong Wang. Statistical methods for identifying differentially expressed genes in rna-seq experiments. *Cell & bioscience*, 2(1):1–8, 2012.

[3] George Michailidis and Florence d'Alché Buc. Autoregressive models for gene regulatory network inference: Sparsity, stability and causality issues. *Mathematical biosciences*, 246(2):326–334, 2013.

Dynamic genes network inference from very short time series.
How repeated acoustic stimuli affect plant immunity?

Khaoula HADJ-AMOR[1] Adelin BARBACCI[2] Frédérick GARCIA[1]

1 Unité de Mathématiques et Informatique Appliquées de Toulouse (MIAT) , Institut National de Recherche pour l'Agriculture, l'alimentation et l'Environnement (INRAe), 24 Chemin de Borde-Rouge, 31326, Castanet-Tolosan, FRANCE

2 Laboratoire des Interactions Plantes Microorganismes et Environnement (LIPME) , Institut National de Recherche pour l'Agriculture, l'alimentation et l'Environnement (INRAe) - Centre National de la Recherche Scientifique (CNRS), 24 Chemin de Borde Rouge, 31326, Castanet-Tolosan, FRANCE

Context and Objectives

RNA-sequencing methods are central to biology and the basis of many breakthroughs in the understanding of the transcriptome. Also, their relatively moderate cost now enables the acquisition of transcriptome time series, which are useful to understand the dynamics of a biological phenomenon after a stimulus. However, **the high dimensionality of RNA-seq data makes it hard to understand the complex dynamics that exists between measured genes**.

Dynamic network inference methods are methods of choice to integrate the complexity of time-dependant transcriptomic data: these methods produce networks in which nodes are genes and an edge between two genes represents a action (activation or repression) of one gene on the other during the time series. However, most network inference methods are also limited by the large dimension of RNA-seq data, in which the number of observations is always much lower than the number of measured gene expressions. Hence, data dimension is frequently reduced by performing the inference on a small number of genes, picked *a priori*. **We propose another strategy, which consists in grouping genes with similar expression dynamics to form a clustered gene network.**

Our strategy is illustrated on an original transcriptomic data associated with the priming of plant immune resistance by repeated acoustic stimuli (RAS).

Dynamic genes network inference from very short time series.

Differential analysis

Differential analysis of genes differentially expressed (DE) by RAS was performed using the R package DESeq2. We considered the gene expression for plants not exposed to RAS as the reference for the analysis. DE genes for plants exposed to 1, 3, 8 RAS were identified by three paired statistical tests. Gene expressions associated with Bonferroni adjusted p-values below 0.05 were considered DE.

Normalization

Counts of DE genes were transformed on log(CPM (count+1)), where CPM is the counts per million.

Classification - sign variation method

For each gene i, we defined the signature vector $s^i = sign(X_{i,t+1} - X_{i,t})$, $t = 1, ..., T-1$. Genes $\{1, \cdots, N\}$ with similar signature were grouped in the same class (2^{T-1} classes). k-means clustering with a different number of clusters was then computed for every class to form c clusters. The silhouette coefficient was computed then for each cluster to determine the number of clusters.

Benchmarking sign variations clustering method:

Method	K	D_{within}	$D_{between}$	DB	Sil	T (s)	R package
Sign variation method	40	0.41	2.30	0.96	0.24	0.03	
k-means for functional data	29	0.60	2.32	1.23	0.20	696.54	fda.usc
Hierarchical for time series data	46	1.29	4.28	1.42	0.23	19.669	dtwclust
K-means longitudinal data	26	0.52	2.31	5.28	0.32	13.653	kml
k-means clustering	28	0.51	2.31	1.09	0.32	0.044	stats
Hierarchical clustering	47	0.45	2.30	6.56	0.25	4.293	stats

Table 1. Benchmark of sign variations clustering method with other clustering methods. K is the number of clusters, D_{within} is the distance within cluster, $D_{between}$ the distance between clusters, DB the Davies-Bouldin index, Sil the average silhouette width, T the computation time (in seconds)

Network inference

First order autoregressive model is written: $X^i_{t+1} = \sum_{j=1}^{n} A_{i,j}\, X^j_t + \epsilon_i$ where $A = (a_{i,j})$ is an $K \times K$ matrix estimated by VAR1 (K is the number of clusters), and ϵ_t is a white Gaussian process. to estimate the coefficient matrix A, we used the the R package SIMONE (Statistical Inference for MOdular Network), which based on a weighted LASSO estimator to take into account information about the network structure. This penalized method leads to the estimation of a sparse coefficient matrix A.

Effect of the clustering method on the network inference:

Method	K	$\#C_{max}$	N_e (% of K^2)	N_{iso} (% of K)	d_{max}	d_{mean}	err
Sign variation method	40	652	36 (2%)	8 (20%)	6	2.25	0.8938
k-means for functional data	29	908	20 (2%)	7 (24%)	7	1.81	0.9380
Hierarchical for time series data	46	1188	32 (1%)	14 (30%)	7	2	0.9494
k-means for longitudinal data	26	629	20 (2%)	6 (23%)	4	2	0.9334
k-means clustering	28	562	32 (4%)	1(3%)	5	2.37	0.9169
Hierarchical clustering	47	970	39 (0.1%)	13 (27%)	7	2.29	0.9358

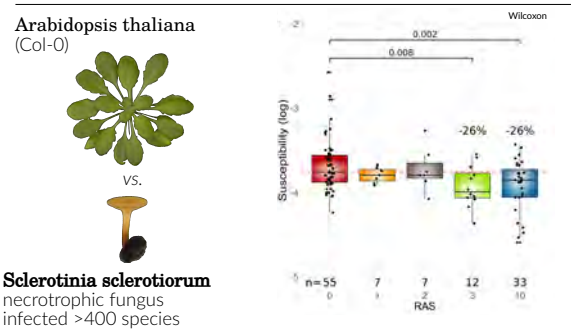
Table 2. Effect of gene clustering method on inferred networks. K is the node number, equal to the number of clusters, $\#C_{max}$ is the maximum cluster size, N_e is the number of edges, N_{iso} is the number of the isolated nodes, d_{max} and d_{mean} are the maximum and mean node degrees, err is the mean square error.

Take home message

- Sign variations clustering is a simple and suitable method for clustering dynamic gene expression of very short time series
- VAR(1) auto-regressive method allow the network inference of clusters computed by the sign variations method

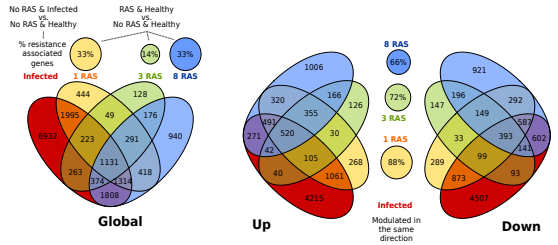
How repeated acoustic stimuli (RAS) affect plant immunity?

Experimental setup and analysis of phenotypes



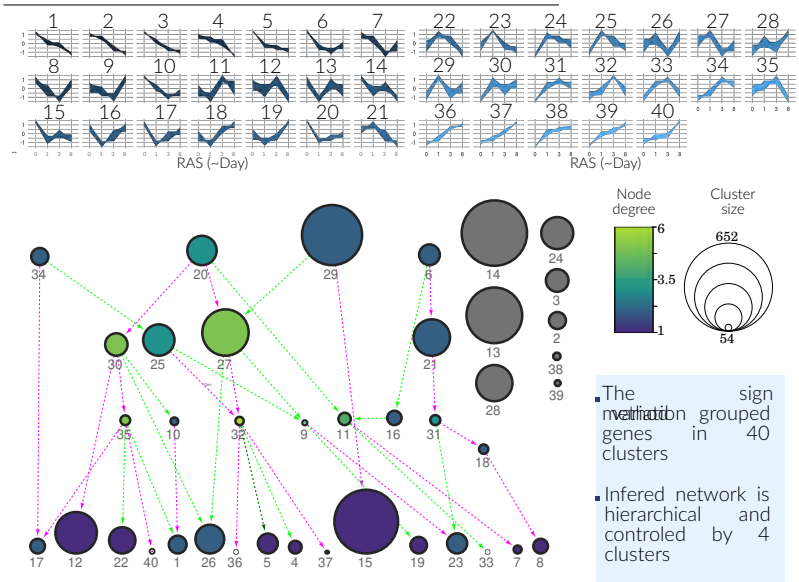
- Plant are exposed to RAS (1KHz, 100dB, 3h/day) during 0 to 10 days prior to infection by the necrotrophic fungs
- A significant 26% gain of resistance (inverse of susceptibility) is observed after 3 RAS

RAS caused the massive transcriptomic reprogramming



- Healthy plants exposed to 1, 3, 8 RAS exhibited a deep transcriptomic reprogramming
- Up to 33% of resistance-associated genes are modulated by RAS in healthy plants

RAS modulated genes are hierarchically regulated



- The sign variation grouped genes in 40 clusters
- Infered network is hierarchical and controled by 4 clusters

Biological conclusions

- Dynamic genes network inference allow the description of the effect of RAS on plant transcriptom. In healthy plants, differentially associated genes are modulated prior to the infection
- The massive transcriptomic reprogramming cascades from 4 RAS-modulated clusters and explain the memory effect

khaoula.hadj-amor@inrae.fr

How repeated acoustic stimuli increase Arabidopsis resistance to the necrotrophic fungus Sclerotinia sclerotiorum?

Khaoula Hadj Amor*[1], Frédérick Garcia[1], Ophélie Léger[2], Mehdi Khafif[2], Sylvain Raffaele[2], and Adelin Barbacci[2]

[1]Université de Toulouse, INRAE, Mathematiques et Informatique Appliquées de Toulouse (MIAT); 31326 Castanet-Tolosan, France. – Université de Toulouse, INRAE, Mathematiques et Informatique Appliquées de Toulouse (MIAT); 31326 Castanet-Tolosan, France. – France

[2]Université de Toulouse, INRAE, CNRS, Laboratoire des Interactions Plantes Micro-organismes Environnement (LIPME); 31326 Castanet-Tolosan, France. – Université de Toulouse, INRAE, CNRS, Laboratoire des Interactions Plantes Micro-organismes Environnement (LIPME); 31326 Castanet-Tolosan, France. – France

Abstract

The repetition of acoustic stimuli (RAS) is particularly effective in priming plant resistance to pathogens. Arabidopsis plants exposed to 3h of sound per day (1KHz, 100dB) exhibited a 12% gain in resistance compared to c.a. 1% gain obtained by genetic engineering. Intriguingly, the resistance gain only occurs after a specific RAS number and remains stable thereafter. The molecular bases associated with this resistance gain remain unknown.
We studied the effect of repeated acoustic stimuli on gene expression. In healthy plants, RAS triggered the expression of c.a. 50% of Arabidopsis resistance-associated genes (> 4000 genes). Plants exhibited extensive transcriptomic reprogramming over increasing RAS involving dominant memory mechanisms. To explore this irreversible dynamic molecular mechanism, we developed dynamic gene network inference methods tailored for very short-time series. The inferred network exhibited hierarchical and robust topology. The inferred network revealed the active role played by the mechanosensitive channel AtMSL10 in RAS priming dynamic implementation.

Keywords: Acoustic, defense priming, Arabidopsis thaliana, Sclerotinia sclerotiorum, plant immunity, dynamic network inference, MSL10

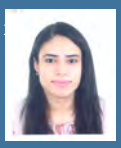
*Speaker

How repeated acoustic stimuli increase *Arabidopsis* resistance to the necrotrophic fungus *Sclerotinia sclerotiorum*?

Khaoula Hadj Amor [1], Frédérick Garcia [1], Ophélie Léger [2], Mehdi Khafif [2], Sylvain Raffaele [2], Adelin Barbacci [1]

1 Unité de Mathématiques et Informatique Appliquées de Toulouse (MIAT) , Institut National de Recherche pour l'Agriculture, l'alimentation et l'Environnement (INRAe), 24 Chemin de Borde-Rouge, 31326, Castanet-Tolosan, FRANCE

2 Laboratoire des Interactions Plantes Microorganismes et Environnement (LIPME) , Institut National de Recherche pour l'Agriculture, l'alimentation et l'Environnement (INRAe) - Centre National de la Recherche Scientifique (CNRS), 24 Chemin de Borde Rouge, 31326, Castanet-Tolosan, FRANCE



Context and Objectives

The repetition of acoustic stimuli (RAS) is particularly effective in priming plant resistance to pathogens. Arabidopsis plants exposed to 3h of sound per day (1KHz, 100dB) exhibited a 12% gain in resistance compared to c.a. 1% gain obtained by genetic engineering. Intriguingly, the resistance gain only occurs after a specific RAS number and remains stable thereafter. **Understanding how acoustic wave-induced vibrations increase plant resistance to pathogens is promising to design sustainable crop health management methods** [ ].The molecular bases associated with this resistance gain remain unknown. We studied the effect of repeated acoustic stimuli on gene expression. In healthy plants, RAS triggered the expression of c.a. 50% of Arabidopsis resistance-associated genes (> 4000 genes). Plants exhibited extensive transcriptomic reprogramming over increasing RAS involving dominant memory mechanisms. To explore this irreversible dynamic molecular mechanism, we developed dynamic gene network inference methods tailored for very short-time series[ ].

How repeated acoustic stimuli (RAS) affect plant immunity?

Experimental setup and analysis of phenotypes

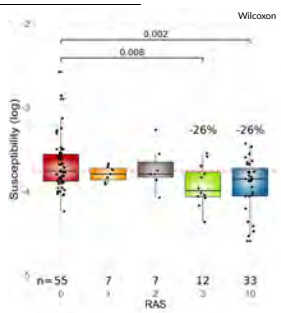
Healthy plants are exposed to RAS (1KHz, 100dB, 3h/day) during 0 to 10 days **prior to infection** by the necrotrophic fungs

A significant **26% gain of resistance** (inverse of susceptibility) **is observed after 3 RAS**

Arabidopsis thaliana (Col-0)
5 week old plants

vs.

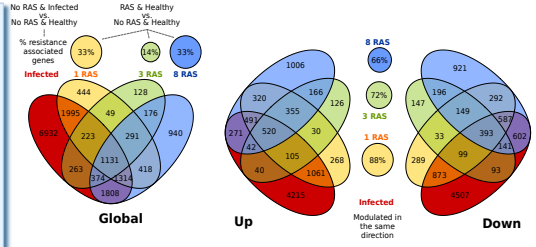
Sclerotinia sclerotiorum
necrotrophic fungus
infecting >400 species



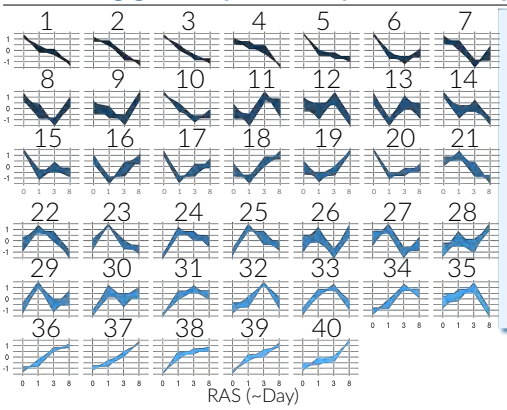
RAS caused the massive transcriptomic reprogramming

Up to 33% of resistance-associated genes are modulated by RAS in healthy plants

RAS-modulated genes depend on the repetition number suggesting **strong memory effect**

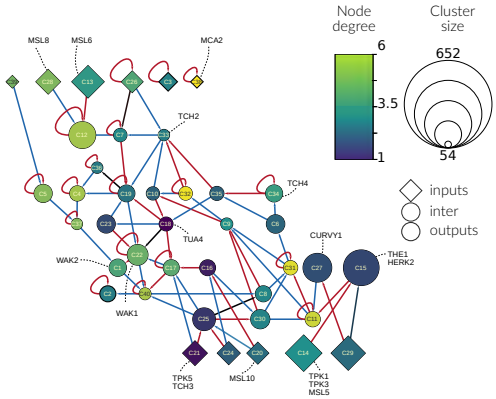


Clustering genes by similar dynamics of expressions



Genes differentially expressed are grouped together by an *adhoc* method we developed for the **clustering of very-short time series unevenly distributed**. **More about the method:**

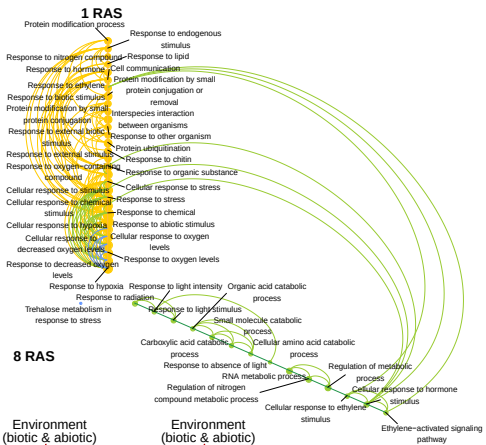
Dynamic RAS-modulated genes expression

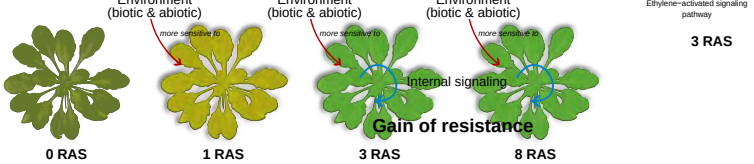


Mechanosensitive channels (MSLs, TPK) are in **input nodes**. They participate in the **regulation of a highly interconnected network**. The network topolgy showed that **genes in a cluster can be modulated by several parent nodes**. It involved that **the priming of plant immunity by RAS poorly depends on a reduced number of RAS-sensitive genes.**

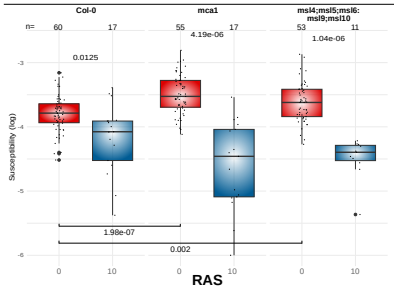
Modeling the effect of RAS on gene expression

Enrichment analysis describes **a progressive variation** of plant perception over RAS. After **1 RAS**, plants are **more sensitive to external signals**, after **3 RAS** plants are **more sensitive to internal signals** associated with hormones





The RAS-associated resistance gain does not depend on a limited number of genes



KO mutants lines mca1 and Δ5 msl exhibited an import loss of resistance compared to the wild type (Col-0). **Resistance phenotypes of impaired mutant lines were rescued by 10 RAS.**

RAS triggered the expression of resistance-associated genes in healthy plants
RAS-increased resistance does not depends on a limited number of genes.
RAS can restore resistance of immunity impaired genotypes