

Thèse de Doctorat

Marc LEGEAY

*Mémoire présenté en vue de l'obtention du
grade de Docteur de l'Université d'Angers
sous le sceau de l'Université Bretagne Loire*

École doctorale : Sciences et technologies de l'information, et mathématiques

Discipline : Informatique et applications, section CNU 27

Unité de recherche : Laboratoire d'étude et de recherche en informatique d'Angers (LERIA)
Institut de recherche en horticulture et semences (IRHS)

Soutenue le 12 décembre 2017

Étude de la régulation anti-sens par l'analyse différentielle de données transcriptomiques dans le domaine végétal

JURY

- Rapporteurs : **M^{me} Céline ROUVEIROL**, Professeure des universités, Université Paris-Nord
M^{me} Fariza TAHI, Maître de conférences - HDR, Université d'Evry – Université Paris-Saclay
- Examinateur : **M. Jérémie BOURDON**, Professeur des universités, Université de Nantes
- Directrice de thèse : **M^{me} Béatrice DUVAL**, Professeure des universités, Université d'Angers
- Co-directeur de thèse : **M. Jean-Pierre RENOU**, Directeur de recherche INRA, INRA d'Angers

Remerciements

Je remercie Céline Rouveiro et Fariza Tahi d'avoir accepté de rapporter cette thèse. Je remercie également Jérémie Bourdon d'avoir accepté d'être examinateur de ma thèse.

Merci à Béatrice Duval, ma directrice de thèse, pour sa patience, son encadrement et surtout pour son soutien tout au long de cette thèse. Merci également à Jean-Pierre Renou, mon co-directeur de thèse, pour ses échanges sur la biologie : grâce à lui, la transcriptomique n'a plus de secrets pour moi (ou presque) !

Je souhaite remercier également l'équipe bio-info de l'IRHS qui m'a tout de suite intégré. Je remercie particulièrement Sandra Pelletier et Sylvain Gaillard qui m'ont bien aidé à comprendre les technologies d'acquisition de données transcriptomiques et leurs traitements. Je remercie également Sébastien Aubourg pour son aide et ses conseils. Merci également à Julie Bourbeillon pour son encadrement en début de thèse. Je remercie le reste de l'équipe de m'avoir supporté pendant les réunions d'équipe hebdomadières : Claudine Landes, Martial Briand et Fabrice Dupuis.

Je remercie également les personnes des autres équipes de l'IRHS. Je remercie plus particulièrement Mathilde Orsel-Baldwin pour ses conseils ainsi que de m'avoir fourni les données et expliquer leur contexte d'acquisition. Je remercie également Jean-Marc Celton pour ses explications sur le fonctionnement des transcrits anti-sens.

Je souhaite remercier mes collègues qui ont partagé mon bureau, qui, eux plus que tous, m'ont vraiment supporté pendant ces années. Je remercie donc mon premier collègue Yi Zhou, avec qui j'ai pu converser sur la culture française ainsi que sur la langue française qui n'est pas aisée à apprendre. Merci à Arthur Chambon, qui nous a rejoint, Yi et moi, dans le bureau par la suite, pour toutes les conversations que nous avons pu avoir sur tous les sujets possibles et imaginables ! Par la suite j'ai déménagé, Vincent Vigneron et Fabien Garreau m'ont donc gentiment accueilli dans leur bureau, puis nous avons accueilli Théo Le Calvar par la suite. Merci à vous pour vos discussions et pour vos échanges lors de la dernière année de thèse. Un remerciement particulier à Pierre Desport, qui n'a jamais partagé mon bureau, mais c'était tout comme !

Je remercie aussi les autres collègues du LERIA qui n'ont pas eu la chance de partager mon bureau, mais avec qui je partage mes repas. Merci à vous, Monsieur Chantrein, pour vos discussions toujours sérieuses mais aussi pour votre sens de l'humour et votre capacité de résistance. Merci également à Benoît Da Mota pour ses conseils et pour les discussions techniques où je dois avouer ne pas toujours tout comprendre. Un remerciement particulier à Laurent Garcia, qui n'est pas du genre rancunier, et avec qui j'ai beaucoup de plaisir à travailler, et avec qui on peut rigoler facilement.

Je remercie également les collègues que je croisais, à l'occasion, dans la salle de pause. Merci à Adrien Goëffon, Matthieu Basseur, Frédéric Saubion, Frédéric Lardeux et Éric Monfroy : grâce à nos discussions autour d'un thé ou d'un café avec des spéculoos, j'ai beaucoup appris sur le fonctionnement de l'université, et j'ai pris note de vos conseils pour l'enseignement.

Je remercie également Gilles Hunault, qui m'a bien aidé avec les méthodes statistiques et avec l'apprentissage du langage R.

Un grand merci également à David Genest et Stéphane Loiseau avec qui j'ai effectué mon stage de master et avec lesquels j'ai continué à collaborer pour l'écriture de papiers en début de thèse. J'avais la chance d'avoir mon premier bureau à côté de celui de David et donc il avait la malchance que je vienne souvent l'embêter ! J'ai pris plaisir à converser avec lui à la fois de sujets de recherches et d'enseignements mais également de plein d'autres sujets, notamment une sombre histoire de chouquettes.

Un remerciement particulier pour nos super secrétaires Catherine Pawlonski et Christine Bardaine, qui elles aussi avaient la chance d'être juste à côté de mon premier bureau. Merci à elles pour leur aide, leur soutien et leur bonne humeur. Un grand merci à Catherine qui est très cliente de mes blagues, que personne d'autre ne doit comprendre vu leur absence de réaction ...

Je remercie également l'équipe technique pour son soutien logistique et pour nos échanges. Merci à Éric Girardeau et Frantz de Germain pour leur disponibilité. Je ne manquerai pas d'envoyer mes remerciements par mail à technique !

Je souhaite remercier les personnes qui m'ont accompagné et soutenu lors de cette période de thèse, mes amis du basket et des Gardiens. Merci pour les parties de jeux de rôles du mardi soir, qui m'ont permis de me vider la tête au moins une fois par semaine ! Merci également à Yann qui est toujours là pour me changer les idées.

Enfin je remercie toute ma famille pour son soutien. Merci à ma maman Édith et mon papa Thierry de m'avoir supporté dans mon choix de faire cette thèse et pour s'être toujours occupé que je ne manque de rien. Merci à mon frère Pierre de m'avoir supporté – tout court – pendant toutes ces années de colocation, et de m'avoir préparé de bons plats ! Je remercie mes grands-parents Joseph et Yvette pour leur soutien, et j'ai une pensée pour Georges qui nous a quitté en 2014 et Simone qui est partie en 2017.

J'en oublie sûrement, je m'en excuse, mais je vous remercie quand même !

Table des matières

Introduction générale	13
1 Expression des gènes	17
1.1 La molécule d'ADN	18
1.2 La transcription	22
1.3 Contrôle post-transcriptionnel	23
1.4 La traduction	26
1.5 Méthodes de mesure de l'expression de gènes	27
1.5.1 Puce à ADN	27
1.5.2 Hybridation comparative	28
2 Réseaux de gènes	31
2.1 Une représentation des interactions : le réseau de gènes	33
2.2 Quelques éléments de théorie des graphes	35
2.3 Méthodes d'inférence de réseaux de gènes	40
2.3.1 Méthodes à base de corrélation	42
2.3.2 Méthodes à base d'information mutuelle	45
2.3.3 Méthodes à base de régression	48
2.4 Générateurs de données	51
2.5 Analyse différentielle de réseaux	53
3 Données d'expression du Pommier	59
3.1 Pommier et puce AryANE	60
3.2 Motivations	60
3.3 Définition des données	63
3.4 Statistiques descriptives	67
3.5 Première étude de réseau sur les données	75
4 Analyse fonctionnelle différentielle	79
4.1 Gene Ontology	80

4.2 Analyse fonctionnelle d'un ensemble de gènes	82
4.2.1 Test d'enrichissement fonctionnel	82
4.2.2 Outils pour le test d'enrichissement fonctionnel	84
4.3 Tests d'enrichissement différentiels	86
4.3.1 Méthode	86
4.3.2 Résultats	89
5 Analyse différentielle de réseaux	95
5.1 Extended Core Network	97
5.1.1 Algorithme	97
5.1.2 Évaluation sur des données simulées	99
5.2 Analyse différentielle de réseaux	101
5.2.1 Gènes AS-impactés et motifs de changement	102
5.2.2 Résultats	107
5.3 Reconfiguration des interactions entre gènes AS-impactés	110
5.3.1 Problème de l'arbre de Steiner minimal	111
5.3.2 Méthode	113
5.3.3 Résultats	115
6 Discussion biologique	117
6.1 Contexte biologique	119
6.2 Interprétation de l'analyse fonctionnelle différentielle	119
6.2.1 Interprétation biologique	121
6.2.2 Évolution de l'expression des couples associés à un terme révélé par les anti-sens	126
6.2.3 Identification de sites de fixation connus	127
6.3 Interprétation de l'analyse différentielle de réseaux	131
6.3.1 Interprétation des résultats de l'analyse différentielle de réseau à l'aide de connaissances biologiques	132
6.3.2 Étude de l'enrichissement des motifs de changement	136
6.3.3 Sous-graphes AS-impactés	139
Conclusion et perspectives	143
Bibliographie	147
A Liste des motifs de changements enrichis de 60DAH	159
B Étude des données à partir de WGCNA	165
C Liste de mes contributions	171

Liste des tableaux

2.1	Méthodes d'inférence de réseaux de gènes.	41
3.1	Liste des échantillons utilisés pour former les expériences H et 60DAH.	65
3.2	Évolutions des couples de sondes.	74
3.3	Répartition des transcrits sens et anti-sens dans les modules.	77
4.1	Liste des termes de la GO slim ‘biological process’.	82
4.2	Dénombrement de transcrits spécifiques à la fois pour le pommier et pour Arabidopsis.	89
4.3	Liste de termes révélés par les anti-sens.	93
6.1	Liste complète des termes enrichis par l’analyse des données sens.	121
6.2	Liste complète des termes révélés par les anti-sens.	123
6.3	Évolutions des couples de sondes des gènes d’intérêt.	126
A	Liste des motifs de changements enrichis de 60DAH.	159
B.1	Répartition des transcrits sens et anti-sens dans les modules du réseau 60DAH.	166
B.2	Termes de l’enrichissement fonctionnel des transcrits des modules de 60DAH.	167
B.3	Termes de l’enrichissement fonctionnel des transcrits des modules de H.	168

Table des figures

1.1	Schématisation du processus de synthèse de protéines dans une cellule eucaryote.	18
1.2	Structure de l'ADN.	19
1.3	Schéma des transformations de l'ADN jusqu'à la protéine.	20
1.4	Capture d'écran d'un génome browser.	21
1.5	Schéma du processus de transcription.	22
1.6	Différents effets de la transcription anti-sens.	25
1.7	Schéma du cycle d'elongation de la traduction.	26
1.8	Étapes du protocole d'utilisation d'une puce à ADN.	28
1.9	Principe d'utilisation de la puce à ADN en hybridation comparative.	29
2.1	Couches représentées par un réseau biologique.	34
2.2	Deux graphes pondérés représentés graphiquement et avec leur matrice d'adjacence. . . .	35
2.3	Différentes topologies de graphes.	36
2.4	Exemples de sous-graphe et de chemins.	37
2.5	Différentes visualisations d'un même graphe.	39
2.6	Utilisation de l'App Cytoscape permettant de dessiner le nœud anti-sens proche de son nœud sens.	40
2.7	Réseau de gènes des données du tutoriel WGCNA.	44
2.8	Déroulement de la suppression des interactions indirectes dans ARACNE en utilisant la DPI.	47
2.9	Visualisation des principales étapes de C3NET.	48
2.10	Procédure de GENIE3.	49
2.11	Définition du programme de régulation <i>RP</i> utilisé par LICORN.	50
2.12	Organisation de la génération des données simulées par GeneNetWeaver.	52
3.1	Pourcentages de loci exprimés et niveaux d'expression moyens.	62
3.2	Ratio d'expression sens/anti-sens dans les différents organes du pommier en fonction de la spécificité d'expression des gènes.	63
3.3	Pourcentage de transcrits correspondants avec des petits-ARN.	64
3.4	Étapes de normalisation des échantillons en utilisant la normalisation par les quantiles. .	66

3.5	Box plots des log-intensités pour chacun des 22 échantillons de l'expérience H.	68
3.6	Boxplots des log-intensités pour chacun des 22 échantillons de l'expérience 60DAH. . . .	69
3.7	Distribution des log-intensités pour chacun des 22 échantillons de l'expérience H.	70
3.8	Distribution des log-intensités pour chacun des 22 échantillons de l'expérience 60DAH. .	71
3.9	Moyennes des log-intensités pour chacun des 22 échantillons des expériences H et 60DAH. .	72
3.10	Nombre de gènes exprimés selon un seuil pour chaque échantillon de H et 60DAH.	73
3.11	Nombre de transcrits sens et anti-sens différemment exprimés entre H et 60DAH selon un seuil.	74
3.12	Dendrogramme des gènes de l'expérience 60DAH.	76
4.1	Extrait de l'ontologie “Biological process”.	81
4.2	Capture d'écran de l'outil AmiGO de l'analyse fonctionnelle pour l'ensemble des trans- crits sens.	84
4.3	Analyse fonctionnelle pour l'ensemble des transcrits sens obtenue grâce à BiNGO. . . .	85
4.4	Méthodologie de l'analyse fonctionnelle différentielle.	87
4.5	Résultat graphique de l'analyse fonctionnelle différentielle H/60DAH.	90
4.6	Mise en valeur du résultat de l'analyse fonctionnelle différentielle.	91
5.1	Étapes de l'inférence de réseaux avec ECN.	100
5.2	Box plots des <i>F</i> -mesures pour C3NET et ECN avec différents taux d'acceptation.	102
5.3	Méthodologie de l'analyse différentielle de réseaux.	103
5.4	Illustrations de l'intégration des données anti-sens dans l'inférence de réseau.	105
5.5	Illustrations d'un <i>graphe de changements</i> et d'un <i>motif de changement</i> dans des réseaux inférés par ECN.	106
5.6	Graphe de changements entre les réseaux de gènes S et SAS inférés avec ECN de l'expé- rience 60DAH.	108
5.7	Extended Core Network avec un taux d'acceptation de 0.05 pour les données sens de 60DAH.	109
5.8	Processus d'analyse de la reconfiguration des interactions des gènes AS-impactés.	114
5.9	Arbre de Steiner d'un sous-graphe AS-impacté de 60DAH.	116
6.1	Méthodologie de l'analyse fonctionnelle différentielle et traitements appliqués afin d'in- terpréter les résultats.	120
6.2	Évolutions des gènes étiquetés par des termes révélés par les anti-sens.	128
6.3	Emplacement putatif de la région promotrice de l'anti-sens.	129
6.4	Résultats de l'analyse des promoteurs des anti-sens associés à la fonction “response to cold” et évoluant négativement entre H et 60DAH.	131
6.5	Méthodologie de l'analyse différentielle de réseaux et traitements appliqués afin d'in- terpréter les résultats.	132

6.6	Méthodologie de la comparaison des réseaux inférés avec l'interactome.	133
6.7	Comparaison des interactions inférées avec les interactions de l'interactome.	135
6.8	Motifs de changement de l'expérience 60DAH.	138
6.9	Arbre de Steiner avec motif de changement.	140
B.1	Correspondance des réseaux H et 60DAH.	169

Introduction générale

Contexte de la thèse

Cette thèse est réalisée grâce au soutien du projet GRIOTE de la région Pays-de-la-Loire. L'objectif de ce projet est de développer des méthodes originales de traitement et d'intégration des données générées par les plate-formes régionales permettant d'accompagner les programmes de recherche en biologie, à la fois dans le domaine de la santé et dans celui du végétal. Ces données de spectrométrie de masse, de séquençage haut débit, de génotypage ou encore de transcriptomique sont dites « omics », elles représentent des vues partielles mais bruitées des systèmes cellulaires étudiés. La gestion, l'analyse et l'intégration de ces données hétérogènes et à très forte volumétrie est un enjeu majeur de la bio-informatique. Un autre objectif du projet GRIOTE est de fédérer la communauté des chercheurs en bio-informatique et d'encourager les collaborations entre les laboratoires de biologie et les laboratoires d'informatique. C'est dans le contexte de ce projet qu'une collaboration entre le Laboratoire d'Étude et de Recherche en Informatique d'Angers (LERIA) et l'Institut de Recherche en Horticulture et Semences (IRHS) est née et que cette thèse a été proposée.

Le LERIA est le laboratoire de recherche en informatique de l'Université d'Angers. Les chercheurs du LERIA développent des recherches fondamentales et appliquées en informatique dans les domaines de l'intelligence artificielle et de l'optimisation. Les recherches qui y sont menées consistent d'une part à développer des modèles et formalismes pour l'accès, la représentation et le traitement des connaissances, et d'autre part à concevoir et mettre en œuvre des algorithmes d'optimisation adaptés à la résolution de problèmes combinatoires. Les travaux du LERIA s'intéressent aussi bien aux aspects fondamentaux qu'aux aspects pratiques et aux développements industriels des recherches.

L'IRHS a été créé à Angers en janvier 2012, sous les tutelles de l'INRA, d'AGROCAMPUS OUEST, et de l'Université d'Angers. Il regroupe la majeure partie des forces de recherche en biologie végétale de la région Pays-de-la-Loire. Ce laboratoire conduit des projets de recherche visant à résoudre des questions de qualité et santé des produits du « végétal spécialisé ». Ses principaux objets d'études sont les rosiers et autres espèces ornementales, les fruits à pépins et légumes, les semences et pathogènes. Le laboratoire développe des approches intégrées en coordonnant les efforts et expertises en génétique, épigénétique, génomique, pathologie, physiologie, écophysiologie, biochimie, modélisation, statistiques et bio-informatique. L'équipe de recherche en bio-informatique de l'IRHS a été créée en septembre 2013

et a pour but de rassembler l'ensemble des forces bio-informatiques des différentes équipes afin de travailler en collaboration avec les autres équipes de l'IRHS, mais aussi de développer les thématiques de recherche en bio-informatique.

Motivations et apports de la thèse

Il existe de nombreuses thématiques de recherche en bio-informatique, allant de l'analyse de séquences jusqu'à l'analyse d'images en passant par de la fouille de données, c'est une discipline de recherche très vaste.

L'avènement de technologies permettant de mesurer, à grande échelle, l'activité cellulaire au niveau du transcriptome, du protéome ou du métabolome a permis l'émergence d'une discipline nommée *biologie des systèmes*. Les études menées dans cette discipline ont pour objectif d'intégrer les informations de ces différents niveaux pour produire une vision systémique des mécanismes de régulation au sein de la cellule. À travers cette modélisation, on peut alors comprendre quels ensembles d'acteurs sont à l'origine de certains phénotypes, étudier comment les actions de ces acteurs sont régulées, chercher quelles dérégulations sont responsables de maladies, ou comprendre quels facteurs environnementaux influencent ces régulations.

Les modèles proposés doivent permettre de représenter les interactions entre les nombreux et différents acteurs du système étudié. Les réseaux sont un formalisme privilégié pour cela, et ils sont utilisés dans de nombreuses approches. Les réseaux de gènes permettent de modéliser les interactions entre les différents gènes dans la cellule dans une condition donnée. Ces réseaux peuvent être étudiés notamment grâce aux mesures de l'expression des gènes fournies par les technologies de séquençage à haut débit.

Dans cette thèse, nous nous intéressons aux réseaux de gènes chez le pommier, avec l'objectif d'intégrer dans ces réseaux les acteurs particuliers que sont les transcrits anti-sens.

L'acide ribonucléique anti-sens est une molécule d'acide ribonucléique (ARN) pour laquelle tout ou partie de sa séquence est complémentaire avec d'autres transcrits. La séquence d'acide désoxyribonucléique (ADN) à partir de laquelle l'ARN anti-sens est transcrit est située sur le brin opposé d'un gène, ce qui explique son nom. Les différentes fonctions de l'ARN anti-sens ne sont pas encore totalement connues mais plusieurs études suggèrent qu'il joue un rôle important dans la régulation de l'expression des gènes.

Nous nous intéressons à la transcription anti-sens chez le pommier pour deux raisons. La première est que le pommier est un des organismes étudiés au sein de l'IRHS, pour lequel plusieurs projets de recherche ont été menés ou sont en cours. La seconde est qu'une étude récente, menée au sein de l'institut, a montré que chez le pommier, une large majorité des gènes codant pour une protéine est concernée par la transcription anti-sens. Dans cette étude [Celton *et al.*, 2014], les auteurs révèlent plusieurs choses intéressantes. Premièrement, le pourcentage d'expression anti-sens est plus élevé que celui observé dans

d'autres organismes : en effet, la transcription anti-sens a été identifiée pour 65% des transcrits sens exprimés dans au moins un organe, alors que dans l'organisme du génome de référence pour le domaine végétal, *Arabidopsis thaliana*, des études identifient la transcription anti-sens pour 30% des transcrits sens exprimés. Deuxièmement, l'expression des transcrits anti-sens est corrélée avec la présence de petits ARN qui agissent sur la régulation de l'expression des gènes. Troisièmement, les niveaux d'expression d'anti-sens varient à la fois selon les organes et les fonctions biologiques des gènes : les anti-sens sont plus exprimés dans les fruits et les graines, et majoritairement pour des fonctions de défense de la cellule.

Lors de cette thèse, afin de mettre en évidence l'impact de l'intégration de données anti-sens dans les réseaux de gènes, nous effectuons une étude exploratoire afin de comparer les informations obtenues grâce à des données « classiques » ne contenant que des transcrits sens, avec les informations obtenues grâce à l'ensemble des données contenant donc les transcrits sens et anti-sens. Nous proposons donc une étude différentielle entre les données sens d'un côté, et les données sens et anti-sens de l'autre.

Nous proposons une étude sur l'ensemble du génome du pommier, avec des données de transcription sens et anti-sens, dans le contexte de la maturation du fruit. Pour mettre en évidence l'impact de la transcription anti-sens, nous proposons une méthode originale d'analyse différentielle de réseaux dans laquelle nous comparons deux réseaux de gènes inférés à partir de deux ensembles d'acteurs : d'un côté les transcrits sens qui sont généralement utilisés dans les études transcriptomiques, et de l'autre les données à la fois sens et anti-sens. En calculant les différences majeures entre ces deux réseaux, nous voulons mettre en valeur les interactions qui sont hautement impactées par les transcrits anti-sens et qui seraient ignorées dans une analyse classique.

Nous proposons d'abord une analyse fonctionnelle différentielle qui nous permet d'identifier quelles sont les fonctions biologiques associées aux transcrits anti-sens. Nous définissons les *termes révélés par les anti-sens*, qui sont les catégories de la Gene Ontology qui apparaissent dans l'analyse contenant les données sens et anti-sens, mais qui ne s'observent pas lors d'une analyse contenant uniquement les données sens. Les termes révélés par les anti-sens permettent de mettre en valeur des fonctions biologiques impactées par les transcrits anti-sens dans le contexte étudié.

Ensuite nous intégrons les données anti-sens dans l'inférence de réseaux de gènes. Il a été observé que les méthodes d'inférence de réseaux de gènes produisent généralement beaucoup de fausses interactions, et plusieurs études ont été faites pour minimiser les interactions redondantes. Une étude propose de n'étudier que le cœur du réseau de gènes en ne calculant pour chaque gène que l'interaction la plus significative avec un autre gène. Cette contrainte étant très restrictive, nous proposons d'étendre le cœur de réseau en considérant pour chaque gène un ensemble de ses « meilleures » interactions, ce qui est réalisé par la méthode d'inférence *Extended Core Network (ECN)*. Avec cette méthode d'inférence, un seuil permet de définir quelles interactions sont significatives, permettant ainsi d'étoffer le cœur de réseau tout en gardant un faible nombre de fausses interactions.

À l'aide de cette nouvelle méthode d'inférence, nous avons développé une analyse différentielle de réseaux qui compare deux cœurs de réseau. Le premier cœur de réseau est constitué uniquement de transcrits sens, alors que le second est constitué de transcrits sens et anti-sens. Nous définissons ainsi

la notion de *gènes AS-impactés* qui décrit les gènes sens dont les interactions dans le cœur de réseau sont fortement impactées par la prise en compte des transcrits anti-sens dans l'inférence du réseau. Afin d'étudier les relations entre les gènes AS-impactés et leurs voisins dans les différents cœurs de réseau, nous définissons les *motifs de changement*. Ces motifs de changement permettent d'identifier l'impact des anti-sens sur les sens. Enfin nous proposons également une méthode afin d'étudier la reconfiguration du cœur de réseau autour des gènes AS-impactés. Cette méthode permet de voir comment des gènes sens pour lesquels on observe une interaction avec une analyse classique en n'utilisant que les données sens, peuvent interagir lorsqu'on intègre les données anti-sens.

Organisation de la thèse

La thèse est organisée en six chapitres. Les deux premiers chapitres forment un état de l'art de la thèse.

Le premier chapitre, « Expression des gènes », décrit les notions biologiques nécessaires à la compréhension du travail, à savoir le mécanisme d'expression des gènes, les différents contrôles post-transcriptionnels et les méthodes d'acquisition de données transcriptomiques.

Le deuxième chapitre, « Réseaux de gènes », présente la modélisation des interactions au sein de la cellule par les réseaux de gènes, les différentes méthodes de construction de réseaux et les méthodes plus récentes d'analyse différentielle de réseaux.

Le troisième chapitre, « Données d'expression du Pommier », présente les données d'expression utilisées lors de cette thèse.

Les deux chapitres suivants présentent les méthodes d'analyses différentielles.

Le chapitre 4 « Analyse fonctionnelle différentielle » définit notre méthode d'analyse fonctionnelle et ses résultats.

Le chapitre 5 « Analyse différentielle de réseaux » définit notre méthode de comparaison de réseaux, ainsi que la méthode d'inférence de réseaux que nous avons proposée.

Dans le chapitre 4 et le chapitre 5, nous présentons les résultats bruts obtenus par nos méthodes sur les données du pommier, c'est-à-dire les termes ou listes de gènes révélés par les traitements que nous avons implémentés.

Dans le dernier chapitre, « Discussion biologique », nous donnons une interprétation biologique de ces résultats, en détaillant quelques traitements bio-informatiques que nous avons réalisés pour analyser les informations obtenues. Nous avons choisi de placer l'analyse des différents résultats dans un chapitre final afin d'y rassembler les connaissances biologiques qui sont nécessaires à leur interprétation.

Enfin la thèse se conclut sur des perspectives de poursuite.

1

Expression des gènes

L'expression des gènes est le produit d'une succession de mécanismes complexes qui permettent la synthèse de protéines à partir de l'ADN : de la transcription à la traduction. Plusieurs contrôles s'effectuent entre la transcription et la traduction permettant ainsi de réguler l'expression des gènes.

D'abord nous présentons la manière dont l'information génétique est stockée dans l'ADN, ensuite nous décrivons le mécanisme de la transcription des gènes, le contrôle post-transcriptionnel, puis le mécanisme de la traduction et enfin nous présentons les méthodes de mesure de la transcription des gènes.

1.1	La molécule d'ADN	18
1.2	La transcription	22
1.3	Contrôle post-transcriptionnel	23
1.4	La traduction	26
1.5	Méthodes de mesure de l'expression de gènes	27
1.5.1	Puce à ADN	27
1.5.2	Hybridation comparative	28

1.1 La molécule d'ADN

Les protéines sont des molécules nécessaires à la vie et ont plusieurs rôles dans la cellule. La synthèse d'une protéine est une succession de processus, qui diffèrent entre les procaryotes et les eucaryotes. Dans les organismes eucaryotes, la synthèse de protéines dans une cellule se décompose en deux processus majeurs schématisés par la figure 1.1 : la transcription puis la traduction [Reece *et al.*, 2011]. Nous allons présenter ici les deux processus en détaillant plus particulièrement celui de la transcription.

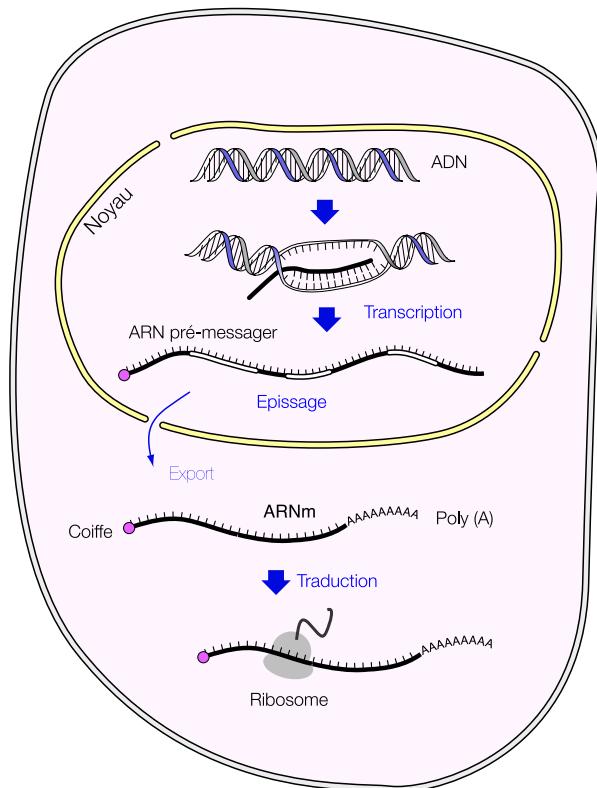


FIGURE 1.1 – Schématisation du processus de synthèse de protéines dans une cellule eucaryote.

Tout commence dans le noyau de la cellule avec l'acide désoxyribonucléique (ADN). L'ADN est une molécule avec une structure en double hélice composée de successions des bases azotées (aussi appelées bases nucléiques) adénine (A), cytosine (C), guanine (G) et thymine (T). La figure 1.2 montre la structure en double hélice de l'ADN ; le lien entre les deux brins de l'ADN s'effectue via les bases azotées, mais une base azotée ne se lie qu'avec un seul autre type de base. Ainsi, en face d'une adénine se trouvera une thymine, et en face d'une guanine se trouvera une cytosine. Cette particularité fait que les deux brins de la molécule d'ADN sont complémentaires. Sur la figure 1.2, en lisant de haut en bas le brin de gauche on obtient ACTG, et le brin de droite, son complémentaire, on obtient TGAC.

L'ADN est la molécule qui contient les informations nécessaires à la synthèse des protéines, appelées informations génétiques. L'ensemble des informations génétiques s'appelle le *génome*. Le génome rassemble l'ensemble des *gènes* de l'espèce : c'est le matériel génétique de l'espèce qui peut être répartie sur un ou plusieurs chromosomes. Le génome est composé de plusieurs régions : des régions codantes,

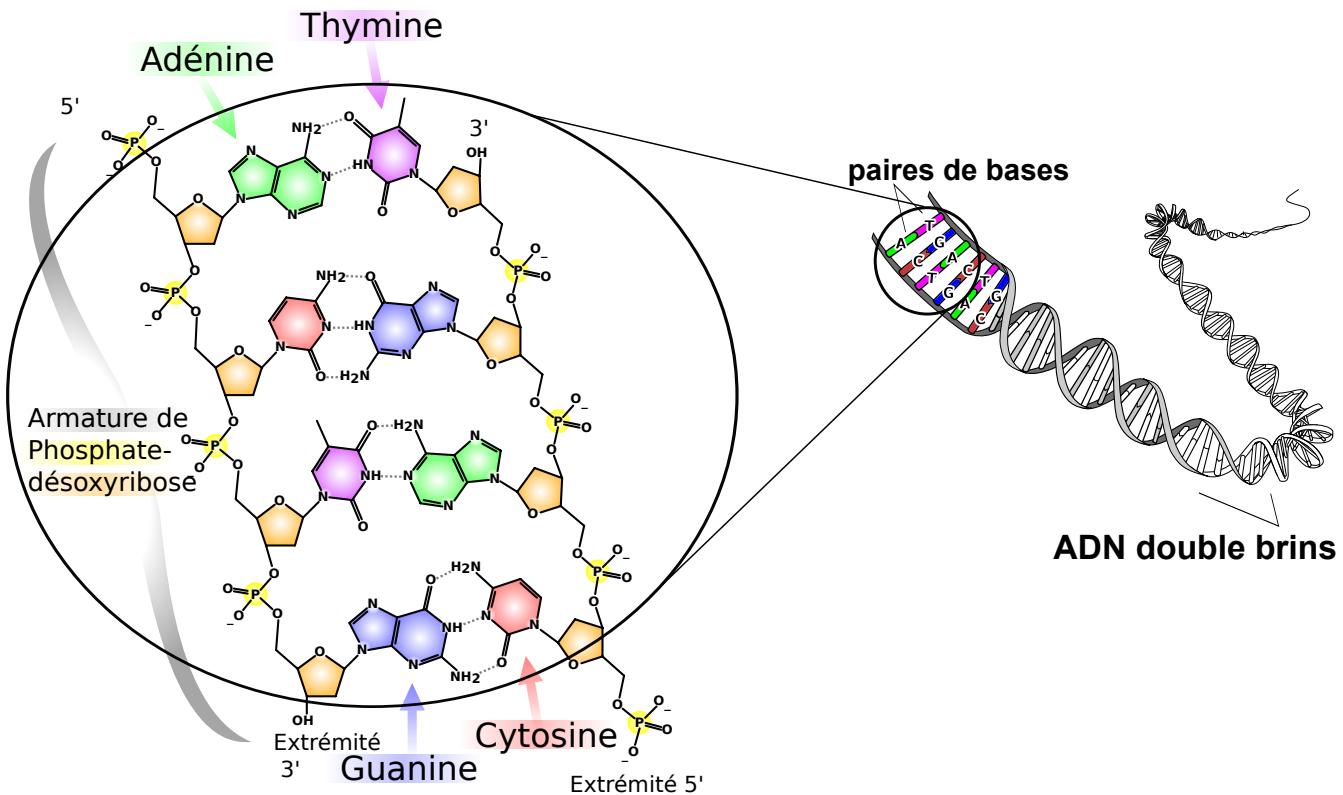


FIGURE 1.2 – Structure de l'ADN.

et des régions non-codantes. Les régions codantes sont les parties de l'ADN qui vont synthétiser une protéine. Le premier schéma de la figure 1.3 montre la structure d'un gène sur un brin de l'ADN. La première extrémité est la région promotrice du gène (en bleu dans l'image), c'est là que se fixent les facteurs de transcription qui activent ou inhibent la transcription. Ensuite le site d'initiation de la transcription (TSS – Transcription Start Site) indique le début du premier exon. Le gène est ensuite une succession d'exons (en vert dans l'image) et d'introns (en jaune dans l'image). Les introns sont les transcrits mais seront éliminés avant la traduction (processus d'épissage). Les exons vont former le transcript mature qui est composé d'une région 5'-UTR (Untranslated Transcribed Region) qui ne sera pas traduite, puis de la séquence codante et enfin d'une région 3'-UTR.

Un gène est situé sur un brin spécifique et sa séquence se lit de la région 5'-UTR vers la région 3'-UTR, c'est ce qui détermine le sens du gène. Les gènes sont eux répartis sur l'ensemble de la molécule d'ADN : ils ne sont pas tous sur un même et unique brin, et ils ne sont donc pas tous dans le même sens. La figure 1.4 est une capture d'écran d'un génome browser, un outil permettant de parcourir un génome. On voit bien sur cette image dans la partie "Named gene" la répartition des gènes sur le chromosome. Les boîtes jaunes indiquent l'emplacement des gènes, ces boîtes ont une pointe à une extrémité qui indique l'emplacement de l'extrémité 3' et donc le sens dans lequel se trouve le gène. Ainsi deux gènes peuvent se superposer en étant sur deux brins différents.

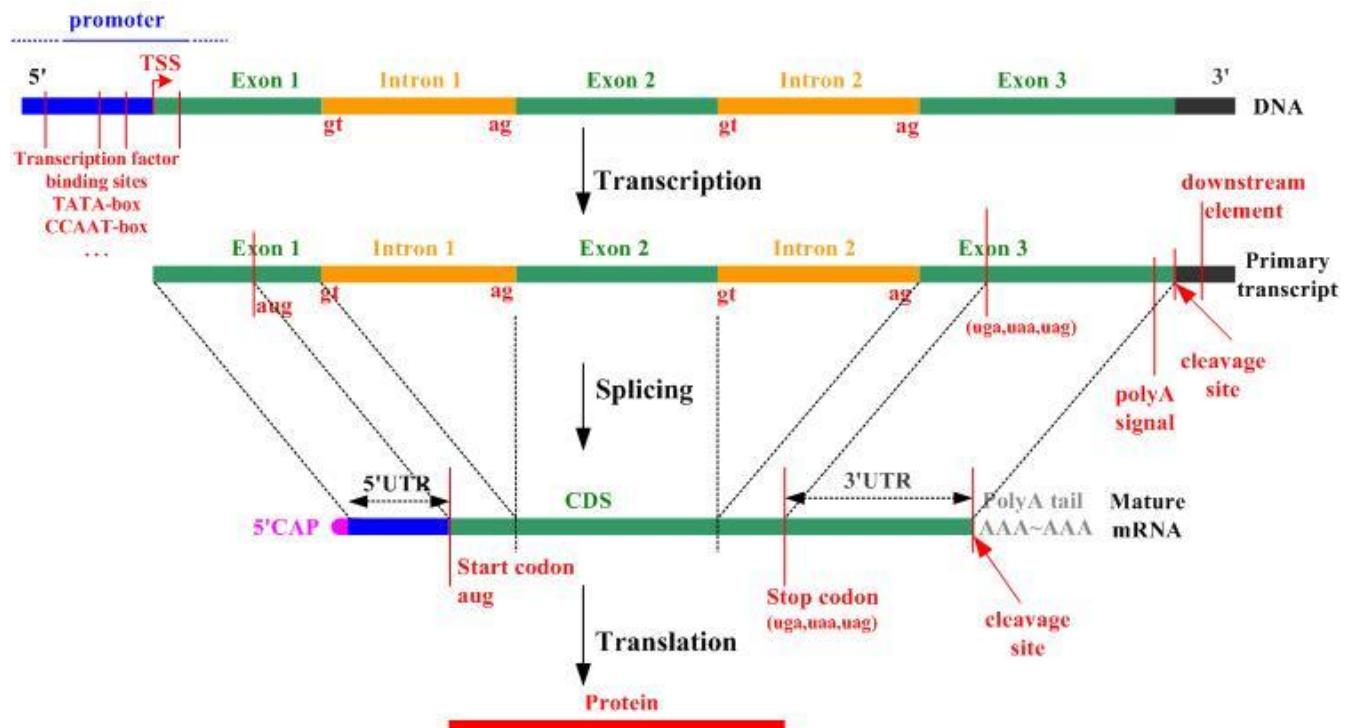


FIGURE 1.3 – Schéma des transformations de l'ADN jusqu'à la protéine. Le schéma indique l'emplacement des différentes régions qui composent un gène sur l'ADN (DNA), puis les régions qui composent un pré-ARNm (Primary transcript) et enfin les régions qui composent l'ARNm mature (Mature mRNA). Le passage de l'ADN au pré-ARN se fait *via* la transcription, puis l'épissage (Splicing) transforme le pré-ARN en ARNm mature, enfin la traduction (Translation) crée la protéine à partir de l'ARNm.

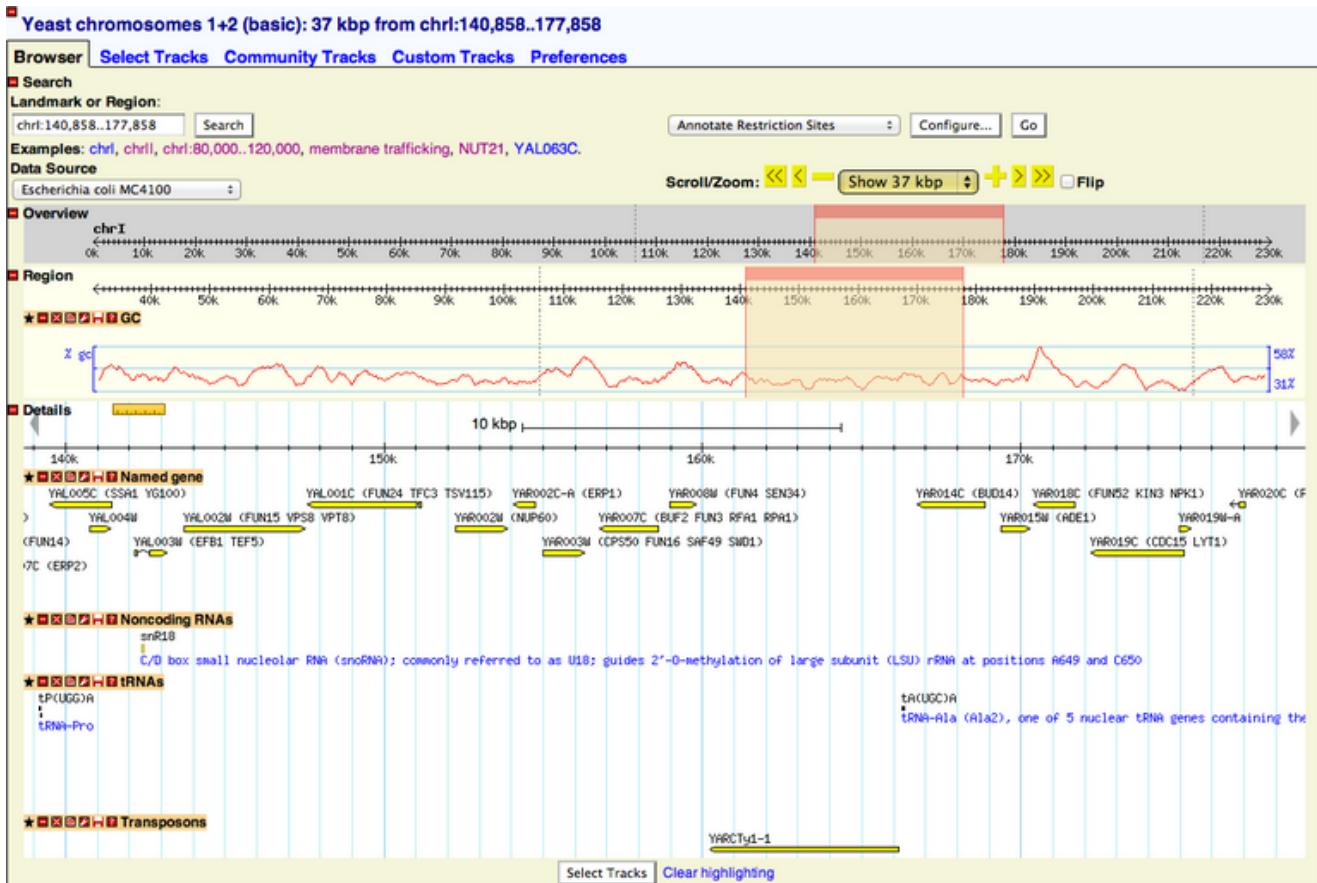


FIGURE 1.4 – Capture d'écran d'un génome browser. La partie haute indique le chromosome actuellement parcouru ainsi que l'emplacement de la fenêtre sur ce chromosome. La partie basse indique l'emplacement de différents acteurs : les gènes (Named gene), les ARN non codant (Noncoding RNAs), les ARN de transport (tRNAs) et des éléments transposables (Transposons).

1.2 La transcription

Le premier processus majeur de l'expression d'un gène est donc la transcription. La figure 1.5 schématise ce processus qui permet de synthétiser l'acide ribonucléique (ARN). Il existe plusieurs types d'ARN qui ont un rôle spécifique ; ici, nous nous intéressons à la synthèse de l'ARN messager (ARNm). La transcription est un processus qui s'effectue en plusieurs étapes dont l'initiation, l'élongation et la maturation.

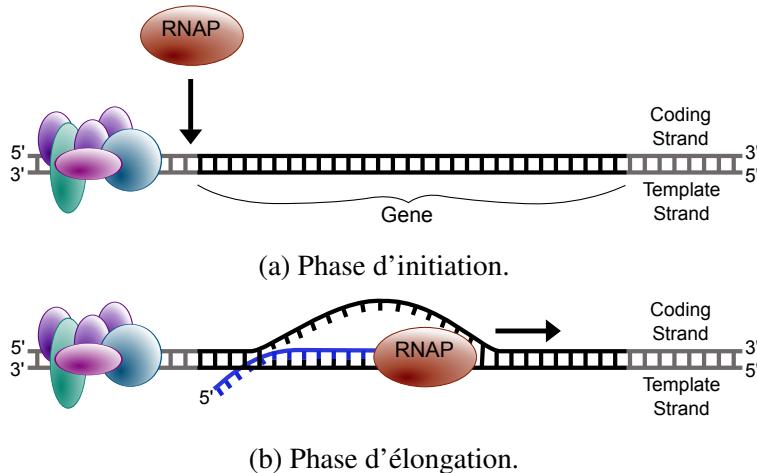


FIGURE 1.5 – Schéma du processus de transcription. (a) L'ARN polymérase (RNAP) repère le promoteur grâce aux facteurs de transcriptions. (b) L'ADN s'ouvre et l'ARN polymérase parcourt un seul brin de l'ADN pour produire l'ARNm (en bleu).

L'ARN polymérase est la protéine qui va parcourir l'ADN afin de produire de l'ARN. En amont de l'UTR 5', un gène comprend des séquences régulatrices et un promoteur, qui indiquent à l'ARN polymérase ("RNAP" dans la figure 1.5) où débuter la transcription. L'étape d'initiation (figure 1.5a) est donc la phase d'approche de l'ARN polymérase à l'ADN. Afin d'initier la transcription, chez les eucaryotes, des protéines appelées des *facteurs de transcription* vont se fixer sur les régions régulatrices, indiquant ainsi à l'ARN polymérase où se trouve le TSS.

L'étape d'élongation (figure 1.5b) est l'étape durant laquelle l'ARN polymérase parcourt l'ADN et assemble les bases nucléiques de l'ARN en fonction des bases nucléiques de l'ADN. L'ARN possède les mêmes bases nucléiques que l'ADN, à l'exception de la thymine remplacée par l'uracile (U) et qui s'assemble également avec l'adénine. L'ARN polymérase va donc se fixer sur un brin de l'ADN et au fur et à mesure que l'ARN polymérase progresse sur l'ADN, il assemble les bases nucléiques de l'ARN par complémentarité de la base lue sur l'ADN. Le brin sur lequel se fixe l'ARN polymérase est appelé brin non-codant ou anti-sens ("Template Strand" dans la figure 1.5). Le brin complémentaire est appelé brin codant ou sens ("Coding Strand" dans la figure 1.5). C'est une convention de nommage qui a été fixée par rapport à la complémentarité des brins et de la transcription. En effet, la séquence d'ARN produite est identique à celle du brin codant, à la seule différence de l'uracile et de la thymine.

L'ARN ainsi produit se nomme l'ARN pré-messager (pré-ARNm), il va subir des modifications lors de la phase de maturation afin de devenir un ARNm. D'abord, une coiffe est fixée sur l'extrémité 5' du

pré-ARNm. Cette coiffe permettra, entre autre, de faciliter la traduction et de protéger l'ARNm d'une dégradation. Ensuite intervient la polyadénylation, processus durant lequel une chaîne d'adénosines (composés chimiques formés à partir d'adénine), appelée queue poly(A), est rajoutée à l'extrémité 3'. Cette queue poly(A) a pour rôle entre autre de guider l'ARNm du noyau vers le cytoplasme et également de protéger l'ARNm d'une dégradation. Enfin le complexe protéique spliceosome va réaliser la phase d'épissage. Lors de cette phase, les introns sont retirés de l'ARN et les exons ainsi rassemblés vont former l'ARNm mature.

1.3 Contrôle post-transcriptionnel

La synthèse d'une protéine est un processus long et complexe. Lorsqu'un organisme est soumis à un changement d'environnement, certains mécanismes doivent s'adapter rapidement afin de survivre à ce changement. Le contrôle post-transcriptionnel est la régulation de la synthèse des protéines par les ARN transcrits. En dégradant ou non les transcrits d'un gène, la cellule peut ainsi réagir plus rapidement au changement d'environnement. Le contrôle post-transcriptionnel intervient ainsi après la transcription et avant la traduction.

Dans la cellule, les ARN transcrits ont une durée de vie limitée, de l'ordre de la minute à quelques heures. Un ARN qui n'est pas traduit sera dégradé et recyclé afin de récupérer les bases nucléiques pour une autre transcription. La dégradation des différents types d'ARN s'effectue par le complexe protéique appelé exosome. L'exosome est présent à la fois dans le noyau et dans le cytoplasme de la cellule, c'est pourquoi, lors de la phase de maturation, l'ARN est protégé contre la dégradation par l'ajout d'une coiffe et d'une queue poly(A), on dit alors que l'ARN est stable. La longueur d'une queue poly(A) détermine la stabilité de l'ARN : plus la queue sera longue plus l'ARN sera stable. Ce « nettoyage » de la cellule fait parti du contrôle post-transcriptionnel en dégradant les ARN qui ont été transcrits et qui, en raison de leur durée de vie, n'ont plus à être traduits.

Un des processus d'inhibition dans le contrôle post-transcriptionnel est l'interférence par ARN [Fire *et al.*, 1998, Baulcombe, 2004]. L'ARN est composé de bases azotées, comme l'ADN. Un brin d'ARN peut donc lui aussi s'hybrider avec un autre brin d'ARN complémentaire formant ainsi un ARN double-brins. L'interférence par ARN est provoquée par la sur-expression d'ARN double-brins et permet de dégrader des séquences spécifiques d'ARN dans le cytoplasme des cellules [Sharp, 2001]. Les ARN double-brins sont détectés par la cellule et sont dégradés. Les produits de cette dégradation sont des petits-ARN dont la taille est comprise entre 21 et 25 nucléotides. Ces petits-ARN vont pouvoir ensuite s'hybrider sur des ARNm et participer à un phénomène nommé l'extinction post-transcriptionnelle (*post-transcriptional gene silencing – PTGS*) [English *et al.*, 1996]. Malgré l'éventuel coût élevé en énergie pour la cellule afin de maintenir ces mécanismes de transcription, ces processus ont été identifiés dans différents organismes, ce qui indique ainsi leur rôle fondamental dans l'adaptation [Berretta et Morillon, 2009, Swiezewski *et al.*, 2009].

ARN anti-sens

Parmi les différents mécanismes impliqués dans le contrôle post-transcriptionnel, nous nous sommes intéressés plus particulièrement au contrôle via les ARN anti-sens. Un ARN anti-sens est une molécule d'ARN endogène dont la totalité ou une partie de sa séquence est complémentaire avec d'autres transcrits [Wang *et al.*, 2005]. C'est un ARN *a priori* non-codant qui est transcrit à partir du brin complémentaire du brin permettant de transcrire l'ARNm. L'ARNm est appelé transcrit sens puisqu'il est le produit de la transcription du brin qui a la même orientation que le gène, alors que l'ARN anti-sens est appelé transcrit anti-sens puisqu'il est le produit de la transcription du brin opposé qui a donc l'orientation inverse au gène. Puisque les transcrits sont issus des brins d'ADN opposés, les séquences régulatrices et promotrices du transcrit anti-sens sont différentes de celles du gène dont il est complémentaire.

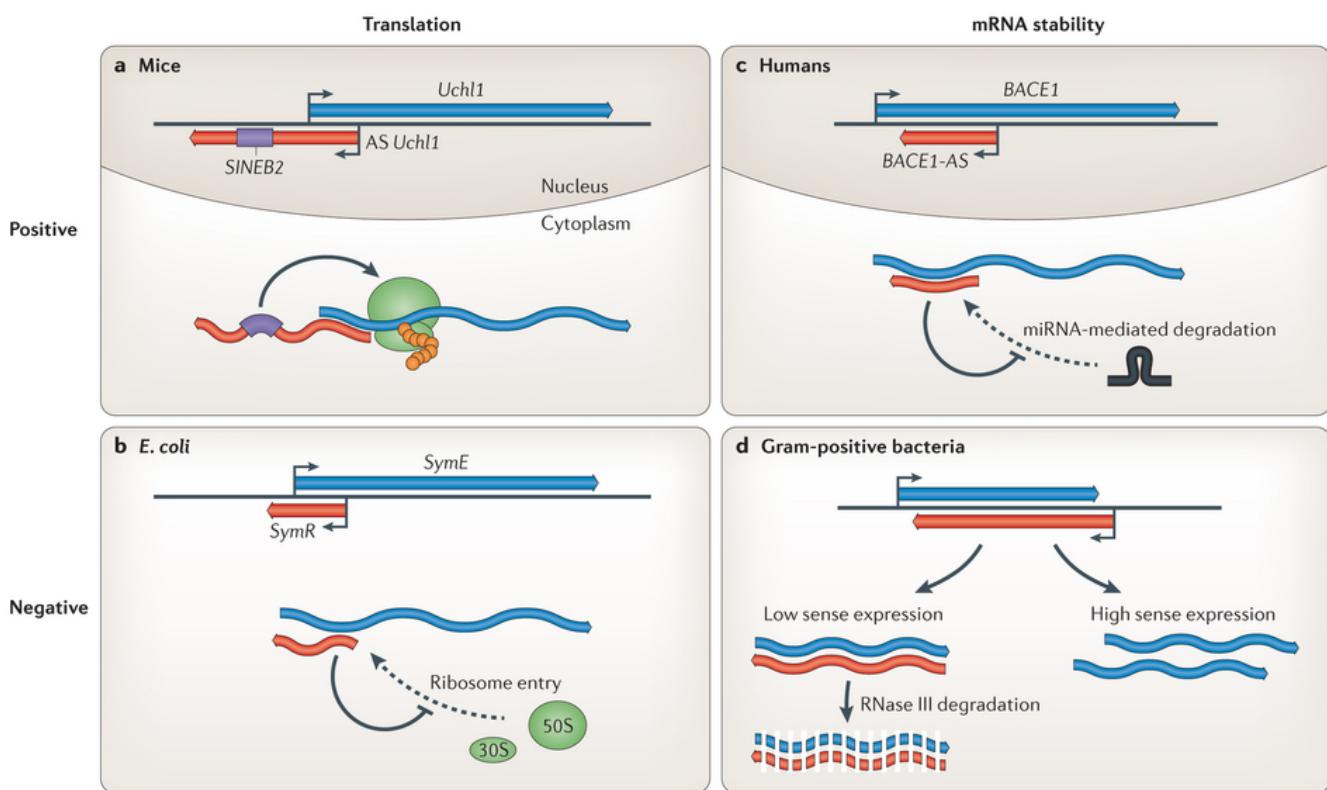
La régulation par l'ARN anti-sens est l'un des mécanismes du contrôle post-transcriptionnel les moins étudiés. Des preuves chez les mammifères, la levure et les plantes suggèrent que la transcription anti-sens joue plusieurs rôles dans la cellule : elle peut avoir un effet positif ou négatif sur l'expression des gènes [Pelechano et Steinmetz, 2013, Murray *et al.*, 2015]. Alors que leur fonction n'est pas totalement connue, les auteurs de [Maruyama *et al.*, 2012] ont identifié des anomalies dans la régulation des transcrits anti-sens dans des tissus dans le cadre d'un cancer du sein, et ils suggèrent ainsi un rôle des transcrits anti-sens dans la formation de tumeurs. Plusieurs études suggèrent également que les transcrits anti-sens peuvent avoir un rôle dans la régulation de la chromatine¹ [Swiezewski *et al.*, 2009, Luo *et al.*, 2013, Murray *et al.*, 2015], et dans l'altération de l'expression du transcrit sens complémentaire dans le phénomène de PTGS *via* la production de petits-ARN [Pelechano et Steinmetz, 2013]. Quelques études effectuées sur les ARN anti-sens [Maruyama *et al.*, 2012, Wang *et al.*, 2014, Celton *et al.*, 2014] ont permis d'identifier qu'au moins 30 à 40% des gènes codants pour une protéine produisent des transcrits anti-sens.

La figure 1.6 illustre différents effets de la transcription anti-sens. La transcription anti-sens peut avoir un effet positif sur la traduction, comme dans la figure 1.6a où chez les souris, la région 5' du transcrit anti-sens du gène Uchl1 s'hybride avec le transcrit sens d'Uchl1 et augmente ainsi l'efficacité de sa traduction. Un autre effet positif de la transcription anti-sens est qu'elle rend plus stable l'ARNm. La figure 1.6c montre cet effet chez les humains où l'hybridation du transcrit anti-sens avec le transcrit sens du gène BACE1 protège ce dernier contre un mécanisme de dégradation en masquant le site de liaison du micro-ARN ("miRNA" sur la figure 1.6) qui régit la dégradation. La transcription anti-sens n'a pas que des effets positifs sur la synthèse des protéines, elle peut avoir un effet de PTGS. La figure 1.6b illustre un phénomène similaire à la figure 1.6c, sauf que cette fois-ci, chez *Escherichia coli*, le transcrit anti-sens SymR du transcrit sens SymE masque le site de liaison de la petite sous-unité de ribosome ("30S" sur la figure 1.6) et empêche donc la traduction de s'initier.

Un autre effet PTGS de la transcription anti-sens est illustré par la figure 1.6d, ici chez les bactéries

¹La chromatine est la structure prise par l'ADN dans le noyau. C'est une association d'ADN, d'ARN et de protéines qui protègent l'ADN.

à Gram-positif. Dans ce cas, le transcrit anti-sens s'hybride en grande partie avec le transcrit sens, ce qui a pour effet de créer un ARN double brin. Cet ARN double brin est ensuite dégradé en étant découpé en petites molécules d'ARN appelées petits ARN interférents (small interfering RNA – siRNA). Ces petits ARN interférents vont ensuite se fixer sur un ARNm cible spécifique et entraîner la dégradation de cet ARNm. L'effet PTGS de ce mécanisme est ainsi double : d'abord l'hybridation du transcrit anti-sens avec le sens puis la dégradation de cet ARN double brin va empêcher la traduction, et ensuite le produit de cette dégradation va lui aussi avoir un effet d'inhibition sur l'expression de gènes, visés par les petits ARN interférents. Comme l'indique la figure 1.6d, cet effet PTGS de la transcription anti-sens s'observe lorsqu'il y a un faible niveau d'expression du sens et de l'anti-sens, mais lorsque l'expression du sens dépasse un certain seuil, cet effet PTGS devient minime et n'empêche plus la traduction de l'ARNm. Le transcrit anti-sens joue donc ici un rôle tampon sur l'expression du sens : lorsque le gène est peu exprimé il sera dégradé par son anti-sens, mais au-delà d'un certain seuil, comme par exemple une situation de stress qui se prolonge, le gène sera bien traduit.



Nature Reviews | Genetics

FIGURE 1.6 – Différents effets de la transcription anti-sens. La transcription anti-sens peut jouer un rôle positif (a et c) ou négatif (b et d) lors de la traduction (a et b) ou dans la conservation de l'ARNm (c et d). Illustration issue de [Pelechano et Steinmetz, 2013].

1.4 La traduction

Lorsque l'ARNm est mature, il est exporté en-dehors du noyau pour le deuxième processus majeur de la synthèse des protéines : la traduction. Une protéine est une molécule formée d'une ou plusieurs chaînes polypeptidiques. Une chaîne polypeptidique est une succession d'acides aminés reliés par une liaison peptidique. Le processus de traduction est le processus durant lequel les acides aminés vont être reliés pour former une chaîne polypeptidique. Les plans de la chaîne sont codés dans l'ARNm : une succession de trois bases nucléiques forme un codon, et chaque codon est associé à un acide aminé. Ce qu'on appelle le code génétique permet de faire la correspondance entre un codon et un acide aminé. Dans le code génétique, on peut remarquer des codons spécifiques : le codon d'initiation et les codons de terminaison. Le premier indique le début de la traduction, et est associé chez les eucaryotes au codon AUG (avec quelques exceptions), les seconds indiquent la fin de la traduction, il s'agit des codons UAA, UAG et UGA.

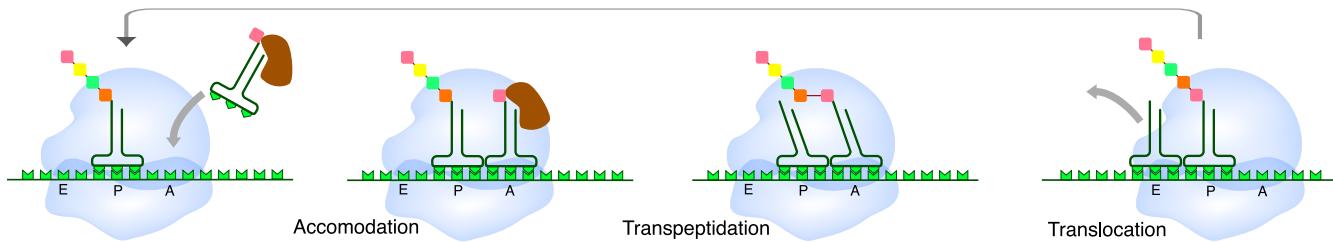


FIGURE 1.7 – Schéma du cycle d'élongation de la traduction. Le ribosome (bleu) parcourt les bases nucléiques (vert clair) de l'ARNm. Lors de l'accommodation, les ARNt (vert foncé) apportent les acides aminés (carrés colorés). Lors de la transpeptidation, l'acide aminé apporté est relié par une liaison peptidique aux autres acides aminés. Lors de la translocation, le ribosome se déplace sur l'ARNm pour lire le prochain codon.

La traduction est un processus complexe qui s'effectue lui aussi en plusieurs étapes. La figure 1.7 schématise une étape de ce processus. Lors de ce processus, l'ARNm est parcouru par un ribosome (en bleu sur la figure 1.7) qui est composé de deux sous-unités : la petite sous-unité (partie inférieure) et la grande sous-unité (partie supérieure). La première étape de la traduction est l'étape d'initiation où la petite sous-unité du ribosome repère le codon d'initiation sur l'ARNm. Un ARN de transport (ARNt) qui porte l'anticodon complémentaire du codon d'initiation s'accroche ainsi à l'ARNm et la grande sous-unité du ribosome vient alors compléter le complexe.

La seconde étape de la traduction est l'élongation, dont la figure 1.7 schématise le cycle. Après que le ribosome soit complet, l'élongation commence son cycle avec l'accommodation. L'accommodation est la phase durant laquelle un ARNt vient se fixer sur le site A du ribosome. Seul un ARNt qui possède un anticodon complémentaire au codon du site A peut se fixer, l'ARNt transporte avec lui un acide aminé correspondant au codon. Une fois que le site A est occupé par un ARNt, l'acide aminé est lié au reste de la chaîne pendant la transpeptidation. Enfin le ribosome se décale d'un codon sur l'ARNm lors de la translocation. Les ARNt étant fixés sur l'ARNm, la translocation décale les ARNt par rapport au

ribosome du site P vers le site E et du site A vers le site P. Cela a pour effet de relâcher l'ARNt maintenant sur le site E, et de libérer le site A pour qu'une nouvelle accommodation puisse se produire. Le cycle de l'elongation se répète jusqu'à arriver à un codon de terminaison.

La terminaison se produit lorsque le ribosome se retrouve avec un codon de terminaison en site A. La chaîne polypeptidique est alors relâchée par le ribosome. Une phase de maturation de la chaîne polypeptidique dans laquelle elle se replie associée à plusieurs modifications biochimiques donnent enfin une propriété fonctionnelle à la protéine.

1.5 Méthodes de mesure de l'expression de gènes

L'expression d'un gène est un succession de phénomènes biologiques comprenant la transcription qui produit de l'ARNm et la traduction qui produit la protéine. On peut mesurer le niveau d'expression d'un gène en mesurant la quantité de transcrits produite par ce gène. En effet, tous les gènes ne produisent pas en continu une quantité égale de transcrits : ils ne s'expriment pas tous de manière uniforme. L'expression d'un gène varie selon la cellule dans laquelle il se trouve et selon les conditions environnementales. Plusieurs méthodes de mesure de l'expression de gènes ont été développées.

Depuis le début des années 2000, le séquençage à haut débit permet de séquencer à moindre coût l'ensemble du génome. Ce séquençage permet d'identifier les séquences des gènes qui sont ensuite utilisées pour mesurer l'expression des gènes. L'une des premières technologies à permettre la mesure d'expression des gènes est l'utilisation de la puce à ADN (microarray en anglais). Une autre technologie utilisée actuellement est le séquençage de l'ARN (RNA-Seq). Nous utilisons dans cette thèse des données produites par des puces à ADN, nous présentons donc dans cette section le fonctionnement de la puce à ADN.

1.5.1 Puce à ADN

Une puce à ADN est constituée d'un ensemble de sondes. Une sonde (ou locus) est un emplacement de la puce qui contient une courte séquence d'ADN simple brin (appelée oligonucléotide), cette séquence d'ADN est connue et représentative du gène dont on étudie l'expression. La particularité de l'ADN simple brin est qu'il reforme sa double hélice en présence d'ADN complémentaire (ADNc) par une réaction appelée l'hybridation. En marquant l'ADNc, il est alors possible d'évaluer le niveau d'expression d'un gène.

La figure 1.8 schématise les étapes du protocole appliqué à une puce à ADN. On extrait tout l'ARN contenu dans un échantillon que l'on met en contact avec des protéines, notamment la rétrotranscriptase qui permet de synthétiser de l'ADNc à partir d'ARN. Un marqueur fluorescent (fluorochrome) est attaché à l'ADNc. Deux fluorochromes sont majoritairement utilisés : la Cyanine 3 (Cy3) et la Cyanine 5 (Cy5) qui fluorescent respectivement dans le vert et dans le rouge. L'ADNc ainsi marqué est placé sur la puce à ADN.

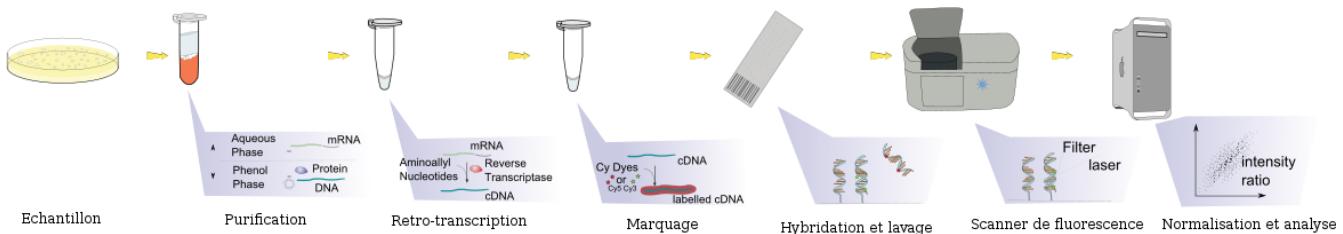


FIGURE 1.8 – Étapes du protocole d'utilisation d'une puce à ADN.

L'ADNc va s'hybrider avec l'ADN simple brin des sondes. La puce est ensuite nettoyée afin de supprimer l'ADNc qui ne s'est pas hybridé puisque reconnu par aucune sonde. Après ce lavage, il ne reste donc que de l'ADN double brin marqué par un fluorochrome, et la quantité d'ADN marqué hybride sur une sonde est proportionnel à l'ADNc ciblé par cette sonde.

L'étape suivante consiste à analyser la puce à ADN à la longueur d'onde d'excitation du fluorochrome. Le fluorochrome ainsi excité va produire une luminescence qui sera enregistrée par le scanner. Cette luminescence est proportionnelle au niveau d'expression du gène correspondant à la sonde. L'image scannée est ensuite analysée pour donner, pour chaque sonde de la puce, un niveau d'intensité lumineuse.

1.5.2 Hybridation comparative

La puce à ADN peut être utilisée pour comparer le niveau d'expression d'un échantillon (appelé traitement) par rapport à un échantillon de référence (appelé contrôle), on parle alors d'hybridation comparative.

Lors d'une hybridation comparative, les deux échantillons sont préparés de manière isolée. Lors de la phase de marquage de l'ADNc, un fluorochrome différent sera affecté à chacun des échantillons (figure 1.9). Ensuite les échantillons sont mélangés sur la puce à ADN, puis le scanner excite l'un puis l'autre des fluorochromes pour ainsi obtenir deux images. L'analyse en sortie du scanner donnera un ratio d'intensité entre les différents fluorochromes, ceci afin de voir si un gène est plus ou moins exprimé dans l'échantillon étudié par rapport à l'échantillon de référence.

Pour éviter un biais technique entre la différence de captation d'un fluorochrome par rapport à un autre, on pratique le dye-swap. Le dye-swap consiste à pratiquer deux analyses et à échanger les fluorochromes. Lors de la première analyse, le traitement est marqué par le Cy3 et le contrôle par le Cy5 par exemple, et donc lors de la seconde analyse, le traitement est marqué par le Cy5 et le contrôle par le Cy3.

L'hybridation comparative est une technique qui est souvent utilisée pour analyser des échantillons « malades » par rapport aux échantillons « sains ». La puce à ADN permet d'identifier rapidement quelles sont les sondes - et donc les gènes - qui ont des niveaux d'expression différents entre le contrôle et le traitement, on parle alors d'analyse différentielle de l'expression.

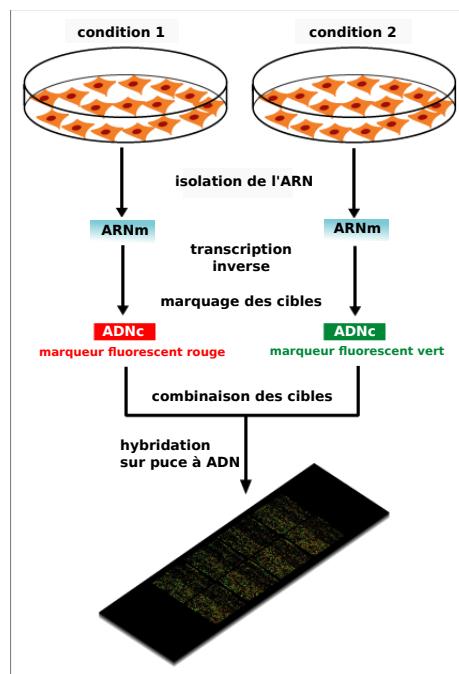


FIGURE 1.9 – Principe d'utilisation de la puce à ADN en hybridation comparative.

2

Réseaux de gènes

Nous avons présenté dans le chapitre précédent quelques uns des nombreux mécanismes de régulation qui opèrent au sein des cellules. Un des objectifs des travaux en bio-informatique est de proposer des modélisations de ces différentes interactions et les réseaux sont un modèle intéressant pour cela.

Notre étude portera sur les réseaux de gènes. C'est pourquoi ce chapitre commence par une brève introduction sur les réseaux de gènes et les interactions qu'ils représentent. Nous rappelons ensuite les propriétés essentielles des graphes qui sont utilisées pour modéliser les réseaux sur lesquels nous allons travailler. La reconstruction de réseaux de gènes à partir de mesures du transcriptome, aussi appelée inférence de réseaux de gènes, a été l'objet de nombreux travaux. Nous présentons quelques méthodes d'inférence représentatives des nombreuses approches disponibles. Enfin, nous expliquons comment une approche récente appelée *analyse différentielle de réseaux* vise à comprendre les mécanismes de régulation ou de dérégulation au sein de la cellule en comparant des réseaux d'interactions associés à des conditions biologiques différentes, comme c'est le cas lorsque l'on compare des échantillons de tissus sains et de tissus malades.

2.1	Une représentation des interactions : le réseau de gènes	33
2.2	Quelques éléments de théorie des graphes	35
2.3	Méthodes d'inférence de réseaux de gènes	40
2.3.1	Méthodes à base de corrélation	42
2.3.2	Méthodes à base d'information mutuelle	45
2.3.3	Méthodes à base de régression	48

2.4 Générateurs de données	51
2.5 Analyse différentielle de réseaux	53

2.1 Une représentation des interactions : le réseau de gènes

L'avènement de technologies permettant de mesurer à grande échelle l'activité cellulaire au niveau du transcriptome, du protéome ou du métabolome ont permis l'émergence d'une discipline nommée *biologie des systèmes*. Les études menées dans cette discipline ont pour objectif d'intégrer les informations de ces différents niveaux pour produire une vision systémique des mécanismes de régulation au sein de la cellule. À travers cette modélisation, on peut alors comprendre quelles régulations sont à l'origine de certains phénotypes, étudier quelles dérégulations sont responsables de maladies, ou comprendre quels facteurs influencent les qualités gustatives d'une pomme en biologie végétale.

Les modèles proposés doivent permettre de représenter les interactions entre les nombreux et différents acteurs du système étudié, et les réseaux sont un modèle intéressant pour cela. Dans un réseau, chaque nœud représente un élément du système biologique comme un métabolite, une protéine ou un gène. Les liens entre chacun des nœuds représentent les interactions entre les éléments du système biologique étudié. Certains réseaux sont spécifiques et ne possèdent qu'un type d'élément : les réseaux métaboliques, les réseaux d'interactions protéine-protéine (PPI), ou les réseaux de gènes. Nous nous intéressons dans cette thèse aux réseaux de gènes.

Il existe des interactions à plusieurs niveaux : au niveau génomique où les gènes sont influencés par leur facteurs de transcription ; au niveau protéomique où les protéines interagissent entre elles et au niveau métabolomique où les métabolites interagissent entre eux et sur les gènes. Toutes ces interactions s'effectuent en permanence et en simultané dans les cellules.

Le réseau de gènes est une modélisation qui projette, sous forme de relations entre les gènes, différents types d'interactions.

Les interactions représentées dans un réseau de gènes peuvent être directes et physiques, comme c'est le cas lorsqu'on examine un facteur de transcription et son gène cible ou lorsqu'il s'agit de deux protéines formant un complexe. Mais une interaction peut aussi être une relation indirecte à travers d'autres éléments qui n'ont pas été mesurés. La figure 2.1, tirée de [Brazhnik *et al.*, 2002], illustre différentes interactions biologiques présentes dans la cellule et comment un réseau de gènes les modélise. Dans cette figure, la protéine 1 produite par le gène 1 interagit avec le gène 2, cette interaction entre le niveau protéique et génomique est représentée dans le réseau de gènes par le lien gène 1 → gène 2 (représenté en pointillés dans la figure). Certaines interactions sont plus complexes : les protéines 3 et 4, respectivement produites par les gènes 3 et 4, forment le Complexe 3-4 qui interagit avec le gène 2 ; cette relation indirecte à travers le complexe protéique est traduite dans le réseau de gènes par les liens gène 3 → gène 2 et gène 4 → gène 2. Enfin les interactions peuvent concerner aussi le niveau métabolomique : le gène 2 produit une protéine 2 qui intervient dans la réaction transformant le métabolite 1 en métabolite 2, le métabolite 2 intervenant dans la transcription du gène 4. Là encore, cette relation complexe est représentée par un lien gène 2 → gène 4 dans le réseau de gènes. Toutes ces interactions sont représentées dans le réseau de gènes par les liens indiqués en pointillés dans la figure.

On voit ainsi que le réseau de gènes est une projection de différents types d'interactions et que des

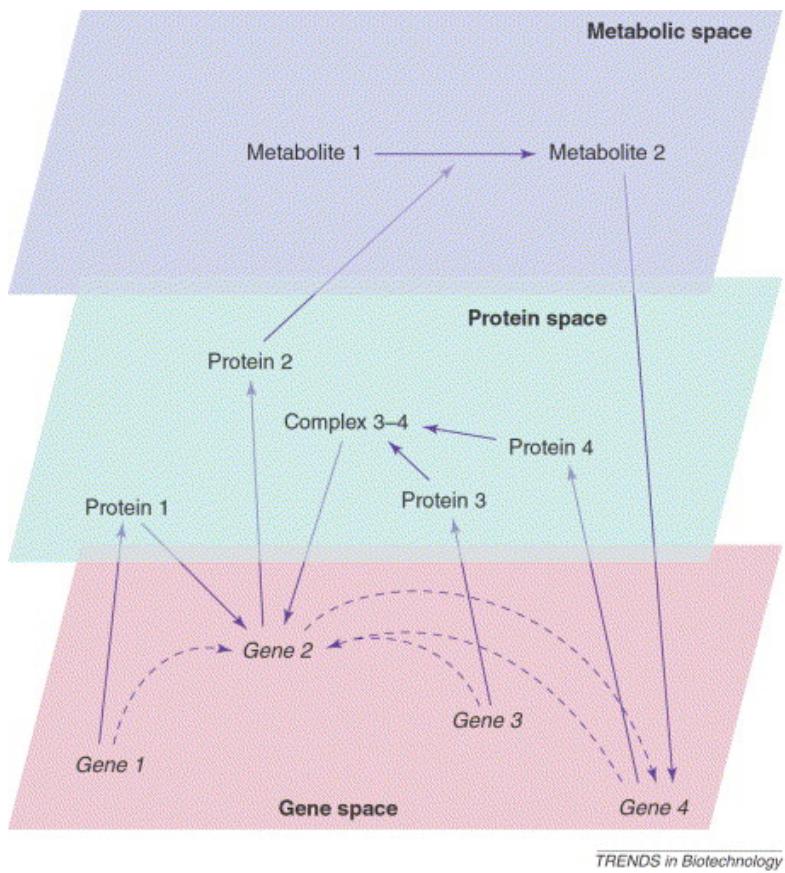


FIGURE 2.1 – Couches représentées par un réseau biologique. Les méthodes d’inférence de réseau modélisent des réseaux de gènes qui font intervenir des relations de plusieurs niveaux biologiques : transcriptomique (Gene space), protéomique (Protein space) et métabolomique (Metabolic space). Illustration issue de [Brazhnik *et al.*, 2002].

acteurs non représentés interviennent indirectement dans ces interactions. Malgré cette simplification et cette agrégation d’informations, les réseaux de gènes sont précieux pour améliorer la compréhension des phénomènes biologiques. Ils permettent d’identifier des modules fonctionnels, c’est-à-dire des ensembles de gènes qui ont de fortes interactions entre eux tout en étant peu soumis aux variations d’autres gènes du réseau. Ils permettent d’identifier des liens entre l’activité de certains gènes et un phénotype donné ou une condition biologique donnée.

L’expression « réseau de gènes », dans la littérature, peut faire référence à plusieurs types de réseaux, en fonction notamment de la nature des interactions représentées. Les *réseaux de régulation* sont des réseaux orientés où une interaction entre deux gènes signifie que la source du lien influe positivement (active) ou négativement (inhibe) la cible du lien. Les *réseaux de co-expression* sont des réseaux non-orientés où une interaction entre deux gènes indique que les profils d’expression des deux gènes sont similaires dans les conditions étudiées.

Dans cette thèse nous nous sommes intéressés aux réseaux de co-expression, l’expression « réseau de gènes » fera donc référence aux réseaux de co-expression. Ces réseaux se représentent à l’aide de graphes présentés dans la section suivante.

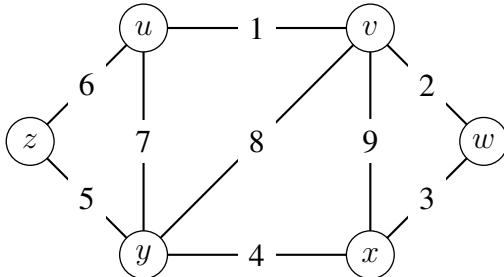
2.2 Quelques éléments de théorie des graphes

Nous rappelons ici les définitions et propriétés essentielles des graphes qui seront utiles dans la suite de notre exposé.

Un graphe $G = (V, A)$ est composé d'un ensemble de nœuds (ou sommets) V et un ensemble d'arêtes A . Les arêtes permettent de relier les sommets entre eux. Le lien formé par l'arête $\{a, b\}$ (notée ab) entre les nœuds a et b est non-orienté : il indique simplement une connexion entre les deux nœuds. Pour représenter un graphe avec des liens orientés, on utilise un graphe orienté $G = (V, A)$, composé de l'ensemble des nœuds (ou sommets) V et de l'ensemble des arcs A . Le lien formé par l'arc (a, b) (pouvant être noté $a \rightarrow b$) du nœud a (appelé source) vers le nœud b (appelé cible) indique la connexion du nœud a « vers » le nœud b .

Le voisinage d'un nœud v est l'ensemble des nœuds v' tel qu'il existe un arc (respectivement une arête) entre v et v' . Le voisinage d'un nœud peut être obtenu grâce à la fonction de voisinage $N : V \rightarrow 2^V$ où $N(v)$ détermine le voisinage du nœud v .

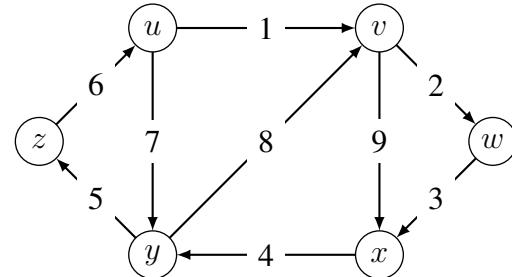
Des étiquettes sont associées aux nœuds ce qui permet de nommer les sommets. On peut également associer une information aux arcs ou arêtes, ce qui permet de spécifier le poids du lien ou bien de spécifier le type du lien : on parle dans ce cas de graphe étiqueté. Lorsque l'étiquette d'un lien représente un poids, on parle alors de graphe pondéré.



(a) Graphe pondéré non-orienté G_{no} .

$$\begin{matrix} & u & v & w & x & y & z \\ u & 0 & \mathbf{1} & 0 & 0 & \mathbf{7} & \mathbf{6} \\ v & \mathbf{1} & 0 & \mathbf{2} & \mathbf{9} & \mathbf{8} & 0 \\ w & 0 & \mathbf{2} & 0 & \mathbf{3} & 0 & 0 \\ x & 0 & \mathbf{9} & \mathbf{3} & 0 & \mathbf{4} & 0 \\ y & \mathbf{7} & \mathbf{8} & 0 & \mathbf{4} & 0 & \mathbf{5} \\ z & \mathbf{6} & 0 & 0 & 0 & \mathbf{5} & 0 \end{matrix}$$

(c) Matrice d'adjacence du graphe G_{no} .



(b) Graphe pondéré orienté G_o .

$$\begin{matrix} & u & v & w & x & y & z \\ u & 0 & \mathbf{1} & 0 & 0 & \mathbf{7} & 0 \\ v & 0 & 0 & \mathbf{2} & \mathbf{9} & 0 & 0 \\ w & 0 & 0 & 0 & \mathbf{3} & 0 & 0 \\ x & 0 & 0 & 0 & 0 & \mathbf{4} & 0 \\ y & 0 & \mathbf{8} & 0 & 0 & 0 & \mathbf{5} \\ z & \mathbf{6} & 0 & 0 & 0 & 0 & 0 \end{matrix}$$

(d) Matrice d'adjacence du graphe G_o .

FIGURE 2.2 – Deux graphes pondérés représentés graphiquement et avec leur matrice d'adjacence. Le graphe G_{no} est non-orienté tandis que G_o est un graphe orienté.

La figure 2.2 représente deux graphes pondérés, un graphe non-orienté $G_{no} = (V, A_{no})$ (2.2a) et un graphe orienté $G_o = (V, A_o)$ (2.2b). Les deux graphes sont composés du même ensemble de nœuds

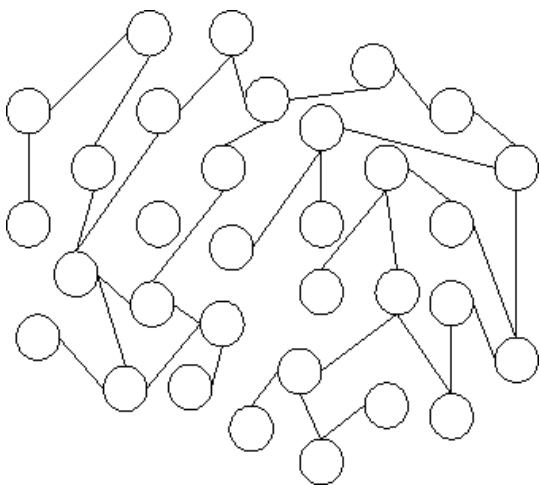
$V = \{u, v, w, x, y, z\}$. Le graphe pondéré orienté est composé de l'ensemble des arcs $A_o = \{((u, v), 1), ((v, w), 2), ((w, x), 3), ((x, y), 4), ((y, z), 5), ((z, u), 6), ((u, y), 7), ((y, v), 8)\}$. Le graphe non-orienté est composé de l'ensemble des arêtes A_{no} construit à partir de l'ensemble des arcs A_o , tel qu'un arc (a, b) est remplacé par une arête $\{a, b\}$.

L'ensemble des liens (arcs ou arêtes) d'un graphe (orienté ou non) peut être représenté par une matrice d'adjacence (figures 2.2c et 2.2d), qui est une matrice carrée étiquetée par les nœuds du graphe. Il existe plusieurs représentations de matrices d'adjacence : lorsque le graphe n'est pas pondéré, la matrice d'adjacence est une matrice booléenne indiquant si un lien existe entre deux nœuds ; lorsque le graphe est pondéré, la matrice d'adjacence peut stocker les poids. Dans le cas d'une matrice d'adjacence pondérée, une valeur spécifique est choisie pour indiquer l'absence de lien. Dans les exemples de la figure 2.2, nous n'avons que des poids strictement positifs et nous avons donc choisi la valeur 0 pour représenter l'absence d'interaction. Lorsque le graphe est non-orienté, la matrice d'adjacence est alors symétrique. Dans le cas d'un graphe orienté, la ligne a de la matrice d'adjacence correspond aux liens dont a est le nœud source, l'interaction $a \rightarrow b$ sera stockée dans la cellule $[a, b]$ de la matrice.

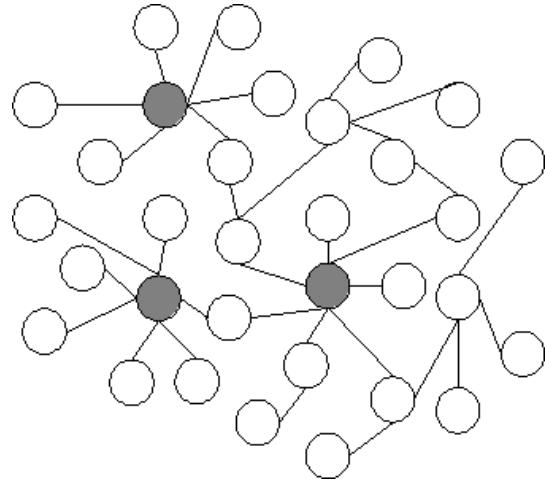
Plusieurs mesures permettent d'étudier les propriétés des graphes et notamment leurs topologies.

Le degré $d(a)$ du nœud a indique le nombre de connexions de a . Dans le cas d'un graphe orienté, on différencie le degré entrant $d^-(a)$ qui est le nombre d'arcs dont le nœud a est la cible, du degré sortant $d^+(a)$ qui est le nombre d'arcs dont le nœud a est la source. Le degré d'un nœud dans un graphe orienté est défini comme la somme des degrés entrant et sortant. Dans la figure 2.2, le nœud x a un degré $d(x) = 3$ pour le graphe non-orienté et $d(x) = d^-(x) + d^+(x) = 2 + 1 = 3$ pour le graphe orienté.

La distribution de degrés $P(d)$ est la probabilité qu'un nœud ait un degré de d . Si un graphe a n nœuds, dont n_d ont un degré de d , alors la distribution de degrés est $P(d) = n_d/n$.



(a) Topologie aléatoire.



(b) Topologie invariante d'échelle.

FIGURE 2.3 – Différentes topologies de graphes.

La distribution de degrés permet de distinguer la topologie des réseaux. La topologie aléatoire (fi-

gure 2.3a) est lorsque la distribution de degrés d'un graphe suit une loi de poisson indiquant qu'approximativement tous les nœuds du graphe ont le même degré. La topologie invariante d'échelle – scale-free en anglais – (figure 2.3b) est lorsque la distribution de degrés d'un graphe suit une loi puissance : $P(d) \sim d^{-\beta}$ où β est la puissance. La puissance β est supérieure à 1 et généralement comprise entre 2 et 3. La plupart des nœuds d'un graphe scale-free ont un très faible degré, et peu de nœuds du graphe, appelés hubs, ont un fort degré et sont des points de connexion importants dans le réseau. Les réseaux biologiques ont généralement une topologie scale-free [Barabási et Albert, 1999, Pržulj, 2007].

Dans des graphes de grande taille, pouvant contenir plusieurs milliers de nœuds, il est intéressant d'identifier comment les nœuds se regroupent et se structurent. En prenant en compte la connectivité des voisins d'un nœud, le coefficient de clustering permet d'évaluer le regroupement des nœuds dans un graphe. Le coefficient local de clustering $C(a)$ d'un nœud a de degré d du graphe $G = (V, A)$ est :

$$C(a) = \frac{2 \times |\{xy \in A | x \in V \wedge y \in V \wedge ax \in A \wedge ay \in A\}|}{d(d-1)} \quad (2.1)$$

$C(a)$ est le ratio entre le nombre de liens existants entre les voisins de a , et le nombre de liens maximum possibles entre ces voisins $(d(d-1)/2)$. Un coefficient de clustering égal à 1 signifie que le nœud appartient à une clique, c'est-à-dire un graphe complet. Si le coefficient est égal à 0, cela signifie qu'il n'existe aucune connexion entre les voisins du nœud. Dans la figure 2.2a, le coefficient de clustering de y est $C_y = \frac{2 \times 3}{4 \times 3} = 0.5$, celui de z est $C_z = \frac{2 \times 1}{2 \times 1} = 1$, z étant effectivement dans une clique avec ses voisins u et y .

Le coefficient de clustering est une mesure intéressante mais ne donne qu'une vue locale sur le rôle d'un nœud. Pour une vue plus globale, on peut considérer des sous-graphes. Un sous-graphe $H = (V_H, A_H)$ de $G = (V, A)$ est défini tel que $V_H \subset V$ et $A_H \subset A$. On dit que H est un sous-graphe de G induit par V_H si $A_H = \{ab \in A | a, b \in V_H\}$. Un sous-graphe de G induit par V_H est donc formé des sommets V_H et de tous les liens existants dans G entre les sommets de V_H . La définition est similaire pour un graphe G orienté. La figure 2.4a montre le sous-graphe du graphe orienté G_o de la figure 2.2b induit par l'ensemble de nœuds $\{v, w, x, y\}$.

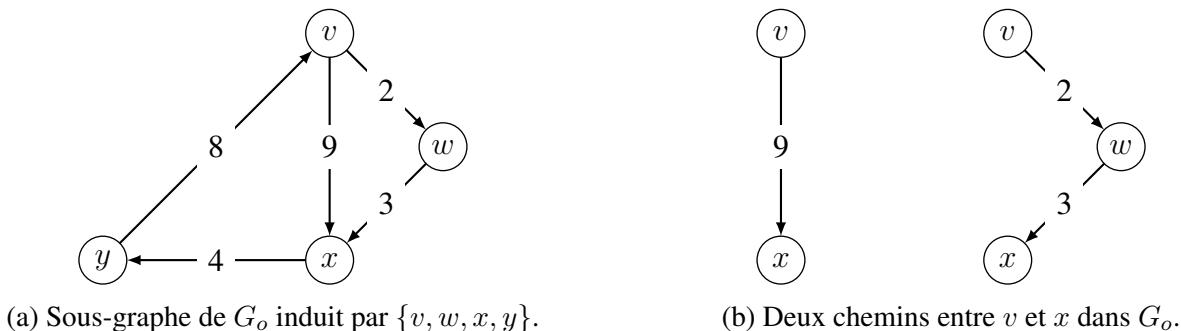


FIGURE 2.4 – Exemples de sous-graphe et de chemins.

Dans un graphe, les distances sont mesurées par la longueur d'un chemin. Un chemin entre a et b est une suite finie de liens consécutifs reliant a à b . La longueur d'un chemin peut être définie comme étant le nombre de liaisons qui séparent deux nœuds, ou, dans le cas d'un graphe pondéré, la somme des poids liant deux nœuds. Puisqu'il peut exister plusieurs chemins entre deux nœuds, la recherche du plus court chemin est un problème pour lequel plusieurs algorithmes ont été proposés en théorie des graphes. Dans le graphe G_o de la figure 2.2a, il existe au moins deux chemins entre les nœuds v et x , représentés dans la figure 2.4b. Le chemin direct $v - x$ entre v et x est le plus court chemin si on ne considère pas les poids, par contre si on considère les poids, le chemin $v - x$ a un poids de 9 tandis que le chemin $v - w - x$ a un poids de $2 + 3 = 5$ et sera donc le plus court chemin.

Visualisation des graphes

Les graphes sont des objets mathématiques dont la représentation graphique permet des visualisations très intéressantes pour leur compréhension et leur exploitation. Ainsi le chemin entre deux points x et y , qui se définit comme une suite de points adjacents dans le graphe, correspond à la notion intuitive de chemin ou d'itinéraire entre x et y si le graphe représente une carte routière par exemple. Dans le domaine de la biologie, les visualisations graphiques des réseaux permettent aux biologistes d'explorer les réseaux étudiés, d'avoir une vue synthétique de leurs propriétés caractéristiques (comme la présence de hubs par exemple), d'effectuer des zooms sur certaines parties jugées intéressantes, *etc.* Bien sûr la taille du graphe induit des limitations certaines sur les visualisations possibles et exploitables.

De nombreux outils existent pour représenter visuellement des graphes. Dans le domaine de la bio-informatique, parmi les outils souvent utilisés, on peut citer Cytoscape [Shannon *et al.*, 2003], Tulip [Auber, 2004] et Gephi [Bastian *et al.*, 2009]. Durant mes travaux, j'ai utilisé essentiellement Cytoscape. Cytoscape est un logiciel de visualisation de réseaux d'interactions moléculaires et de voies biologiques qui permet entre autre d'annoter les réseaux et d'y intégrer des données d'expression de gènes. Cet outil est assez lourd à mettre en œuvre et à maîtriser, mais il offre de nombreuses possibilités. De plus, un avantage de Cytoscape est qu'il propose un ensemble d'Apps [Saito *et al.*, 2012] développés par la communauté qui permettent d'adapter l'outil à des besoins spécifiques.

Un problème essentiel pour la visualisation d'un graphe est de placer les sommets de manière à ce que l'ensemble des interactions soit « lisible ». Cytoscape, comme les autres outils de visualisation, propose différents algorithmes de placement des nœuds ; les nœuds peuvent être placés en grille, en cercle, de manière hiérarchique, ou leur position peut être déterminée par un algorithme qui tient compte de la force des interactions entre les nœuds. Dans ce dernier algorithme, dit dirigé par les forces, les nœuds sont placés en fonction de leur degré, et, si le réseau est pondéré, en fonction du poids des liaisons : les nœuds se repoussent entre eux comme des aimants, une liaison entre deux nœuds va rapprocher les nœuds comme le ferait un ressort, plus le poids de la liaison est fort, plus les nœuds seront proches. Grâce à cet algorithme les nœuds très connectés sont placés au centre du dessin et les nœuds peu connectés sont vers la périphérie du dessin.

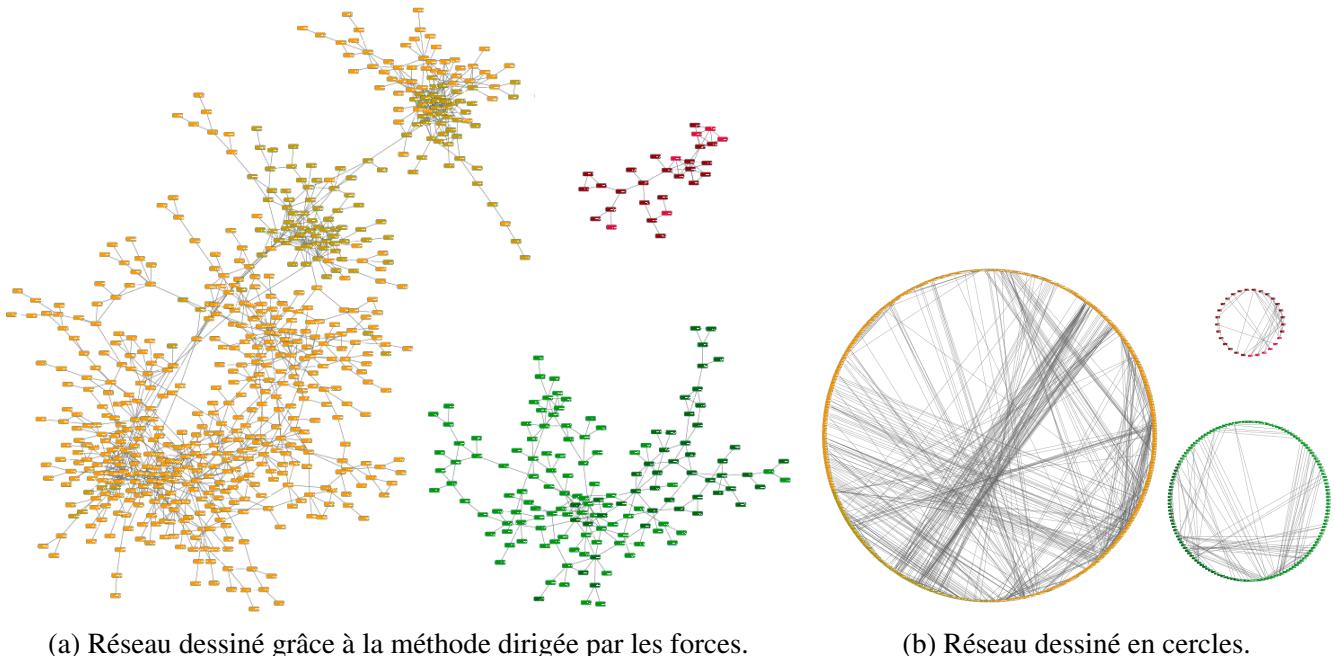


FIGURE 2.5 – Différentes visualisations d'un même graphe.

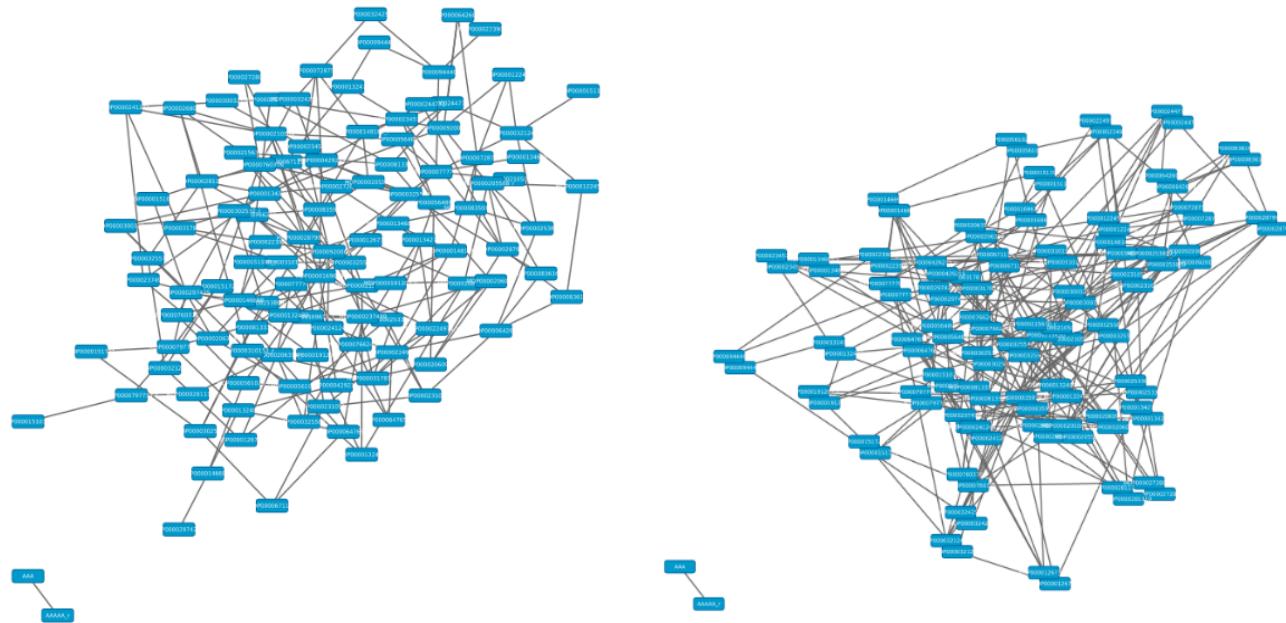
La figure 2.5 montre la visualisation sous Cytoscape d'un même réseau avec deux algorithmes de placement différents. Le placement par la méthode dirigée par les forces (2.5a) permet de visualiser rapidement les modules d'un graphe. Dans un réseau de gènes où un nœud représente un gène, il est alors aisément d'observer les gènes qui sont fortement connectés. Le placement circulaire (2.5b) permet d'identifier les composantes connexes d'un graphe. Une composante connexe est un sous-graphe pour lequel il existe un chemin entre chacun des nœuds du sous-graphe, dans le cas d'un graphe orienté, on ne prend pas en compte l'orientation des arcs. Avec la représentation circulaire, on peut observer les interactions entre chacun des nœuds d'une composante connexe.

Pour la représentation des réseaux de gènes, nous préférons la méthode dirigée par les forces qui a l'avantage ici d'observer la force d'interaction entre deux gènes.

J'ai pu encadrer le développement de deux Apps Cytoscape pour le dessin de réseaux de gènes au cours de ma thèse. Ces développements ont été réalisés par des étudiants de Master Informatique.

La première application concerne la représentation d'un réseau d'interactions pondéré. Dans un tel réseau, chaque interaction est affectée d'un poids, qui est un nombre réel. L'application permet à l'utilisateur de déterminer, de manière interactive, le seuil qui définit si un poids est suffisamment important pour que l'interaction figure dans le réseau et soit dessinée par Cytoscape. Le fait de pouvoir modifier ce seuil permet de visualiser directement l'effet de ce seuil sur la topologie du graphe. Cette application n'est pas distribuée publiquement mais est actuellement utilisée à l'IRHS.

La seconde application développée permet de dessiner un nœud représentant un transcript anti-sens à côté du nœud représentant son transcript sens. La figure 2.6 montre l'application de cet algorithme de pla-



(a) Réseau composé de couples de transcrits sens et anti-sens dessiné avec la méthode dirigée par les forces.

(b) Réseau composé de couples de transcrits sens et anti-sens dessiné avec la méthode dirigée par les forces pour le sous-graphe des transcrits sens, puis les anti-sens sont dessinés à côté de leur transcrit sens complémentaire.

FIGURE 2.6 – Utilisation de l’App Cytoscape permettant de dessiner le nœud anti-sens proche de son nœud sens.

cement sur un réseau de gènes. Dans le but de représenter un réseau contenant des sens et des anti-sens, nous avons développé une application qui permet de forcer le placement des nœuds afin qu'un anti-sens soit placé à proximité de son sens. Ainsi, dans la figure 2.6a on observe bien les interactions entre les différents transcrits, mais on ne voit pas les interactions entre un sens et son anti-sens. En dessinant d'abord le sous-graphe des nœuds sens, puis en dessinant les nœuds anti-sens proches de leur sens complémentaire (figure 2.6b), on observe alors les interactions entre les transcrits ainsi que les interactions au sein d'un couple sens/anti-sens. Dans cet exemple nous n'avions que des couples de transcrits sens et anti-sens, mais il est possible qu'un anti-sens ne possède pas son transcrit complémentaire dans le réseau et il est alors dessiné au même moment que le sous-graphe de transcrits sens. De cette manière, l'anti-sens « orphelin » est dessiné dans le graphe selon la méthode dirigée par les forces et n'est pas placé de manière arbitraire dans le dessin du graphe.

2.3 Méthodes d'inférence de réseaux de gènes

Nous avons rappelé comment les réseaux de gènes représentent différents types d'interaction au sein des cellules. L'*inférence de réseau de gènes* désigne le processus qui consiste à construire des réseaux de gènes à partir de données d'expression des gènes. Les technologies d'analyse du transcriptome, comme les puces à ADN, permettent désormais de mesurer l'ensemble des transcrits exprimés donnant ainsi une

mesure de l'activité de tout le génome, dans une condition donnée. En étudiant un ensemble d'échantillons homogènes, c'est-à-dire où les cellules de différents individus sont placées dans les mêmes conditions, on cherche à inférer des relations de dépendances entre les gènes à partir des mesures de leurs expressions. Nous ne parlerons ici que des approches qui considèrent des données en état stationnaire.

Ce problème, aussi appelé *reverse engineering*, est difficile pour plusieurs raisons. La première est que les données d'entrée sont les niveaux d'expression de n gènes pour p échantillons. Avec les progrès de la technique, le séquençage de l'ensemble d'un génome est possible et n est donc de l'ordre de quelques milliers ($\approx 60\,000$ pour le pommier), alors que l'on ne dispose généralement que de quelques dizaines à quelques centaines d'échantillons ($n >> p$). La seconde raison est que les relations représentées dans un réseau de gènes résultent d'interactions à différents niveaux et qui font intervenir des acteurs que l'on ne mesure pas dans les études de transcriptome.

Depuis une quinzaine d'années, un grand nombre de méthodes ont été proposées pour l'inférence de réseaux de gènes. Des synthèses sur ces méthodes peuvent être trouvées dans [Bansal *et al.*, 2007, Friedel *et al.*, 2012, Emmert-Streib *et al.*, 2012, Marbach *et al.*, 2012].

TABLE 2.1 – Méthodes d'inférence de réseaux de gènes. Les méthodes en gras sont développées plus en détail dans le manuscrit.

Méthode	Modèle	Citation
WGNA	Corrélation	[Langfelder et Horvath, 2008]
Relevance Network	Information mutuelle	[Butte et Kohane, 2000]
ARACNE	Information mutuelle et Data Processing Inequality	[Margolin <i>et al.</i> , 2006]
CLR	Information mutuelle	[Faith <i>et al.</i> , 2007]
MRNET	Information mutuelle	[Meyer <i>et al.</i> , 2007]
C3NET	Information mutuelle	[Altay et Emmert-Streib, 2010]
NARROMI	Information mutuelle	[Zhang <i>et al.</i> , 2013]
GENIE3	Arbres de régression	[Huynh-Thu <i>et al.</i> , 2010]
TIGRESS	Régression et Least Angle Regression	[Haury <i>et al.</i> , 2012]
LICORN	Recherche de motifs fréquents et modèle de régulations coopératives	[Elati <i>et al.</i> , 2007]
H-LICORN	Recherche de motifs fréquents et régression	[Chebil <i>et al.</i> , 2014]
BoolNet	Réseau booléen	[Müssel <i>et al.</i> , 2010]
BNArray	Réseau bayésien	[Chen <i>et al.</i> , 2006]

Dans les différentes approches proposées, on peut distinguer plusieurs familles de méthodes. De nombreux algorithmes reposent sur une méthode à base d'associations locales : il s'agit, par une mesure qui peut être la corrélation ou l'information mutuelle, d'évaluer la dépendance entre chaque paire de gènes, puis, grâce à un seuillage sur la mesure de dépendance, de construire le réseau. Une autre famille d'algorithmes concerne les méthodes à base de modèles graphiques, comme les modèles graphiques gaussiens et les réseaux bayésiens. Enfin plusieurs méthodes utilisent la régression linéaire pour expliquer l'expression d'un gène en fonction des autres. Ces approches sont particulièrement adaptées lorsqu'on travaille avec deux types d'acteurs, les facteurs de transcription d'une part, et leurs gènes cibles d'autre

part.

La table 2.1 liste de manière non-exhaustive des méthodes d’inférences de réseaux. Un ensemble plus important de méthodes est décrit dans [Marbach *et al.*, 2012], où dans le cadre du projet DREAM (Dialogue on Reverse Engineering Assessment and Methods), les auteurs ont testé une trentaine d’algorithmes. Nous présentons ici seulement quelques méthodes qui relèvent, pour la plupart, de la première famille : elles construisent un réseau de dépendances entre les paires de gènes à l’aide de la corrélation ou de l’information mutuelle.

2.3.1 Méthodes à base de corrélation

Pour calculer la corrélation entre chaque paire de gènes, ces méthodes peuvent employer la corrélation de Pearson (équation 2.2) ou la corrélation de Spearman (équation 2.3). La corrélation entre deux gènes g et s , dans p expériences, est définie par :

$$cor(g, s) = \frac{\sum_{i=1}^p (g_i - \bar{g}) \cdot (s_i - \bar{s})}{\sqrt{\sum_{i=1}^p (g_i - \bar{g})^2} \cdot \sqrt{\sum_{i=1}^p (s_i - \bar{s})^2}} \quad (2.2)$$

$$cor(g, s) = 1 - \frac{\sum_{i=1}^p (g_i - s_i)^2}{p(p^2 - 1)} \quad (2.3)$$

Ces mesures de corrélation ne gèrent que les dépendances linéaires entre les vecteurs d’expression. La matrice de corrélation étant de taille $n \times n$, le stockage de cette matrice nécessite un espace mémoire conséquent.

Weighted Gene Co-expression Network Analysis (WGCNA) [Langfelder et Horvath, 2008] est un logiciel basé sur la corrélation et il a de plus pour objectif d’analyser la topologie du réseau obtenu. Le logiciel, implémenté dans un paquet R, permet, entre autres, de reconstruire un réseau de gènes, puis de regrouper les gènes en « modules » ; un module est un ensemble de gènes densément interconnectés.

Afin de pouvoir construire un réseau de gènes *via* une matrice d’adjacence, WGCNA calcule une matrice de corrélation non-signée. La mesure de corrélation utilisée est la valeur absolue de la corrélation linéaire de Bravais-Pearson (équation 2.2), on considère ainsi que deux gènes dont les valeurs d’expression sont corrélées ou anti-corrélées sont des gènes co-exprimés. WGCNA propose ensuite deux types de réseaux : soit le réseau non-pondéré où un lien indique simplement une co-expression entre deux gènes, soit le réseau pondéré où le poids du lien témoigne de l’importance de la co-expression entre deux gènes. Pour construire un réseau non-pondéré, WGCNA propose la méthode du « hard threshold » : un seuil τ est fixé et si la corrélation ne dépasse pas ce seuil, il n’y a pas d’adjacence. Pour le réseau pondéré, WGCNA propose la méthode du « soft threshold » [Zhang et Horvath, 2005, Horvath et Dong, 2008]. Afin de réduire l’impact des faibles corrélations, et d’augmenter celui des fortes corrélations, une fonction puissance est appliquée sur la valeur de corrélation. Ainsi, à l’aide d’un paramètre $\beta \geq 1$, la valeur d’adjacence entre deux gènes g et s est calculée grâce à l’équation 2.4.

$$a_{gs} = |cor(g, s)|^\beta \quad (2.4)$$

Comme on l'a dit précédemment, il a été observé que les réseaux en biologie correspondent souvent à des graphes à topologie invariante d'échelle (scale free), WGCNA cherche à obtenir une telle topologie et propose donc de déterminer la valeur du seuil τ de la méthode « hard threshold » ou la valeur du paramètre β de la méthode « soft threshold » afin de tendre vers cet objectif. Si la distribution de degrés $p(d)$ suit une loi puissance, alors les valeurs $\log(p(d))$ dépendent linéairement de $\log(d)$. WGCNA utilise donc le coefficient de détermination R^2 de la régression linéaire pour déterminer si $p(d)$ suit une loi de puissance. Après avoir déterminé la matrice de corrélation, WGCNA calcule le R^2 associé à chacun des seuils possibles, et laisse ensuite le choix à l'utilisateur de la valeur qu'il souhaite utiliser. Plus la valeur de R^2 sera proche de 1 et plus la topologie du réseau sera scale-free.

Après avoir construit le réseau, WGCNA propose d'analyser le réseau en regardant s'il existe des « modules » de gènes. WGCNA définit un module comme un ensemble de gènes fortement interconnectés. Pour détecter les modules, WGCNA calcule d'abord la *topological overlap matrix (TOM)*. La mesure “topological overlap” entre deux noeuds se calcule de la manière suivante :

$$\omega_{gs} = \frac{l_{gs} + a_{gs}}{\min(k_g, k_s) + 1 - a_{gs}} \quad (2.5)$$

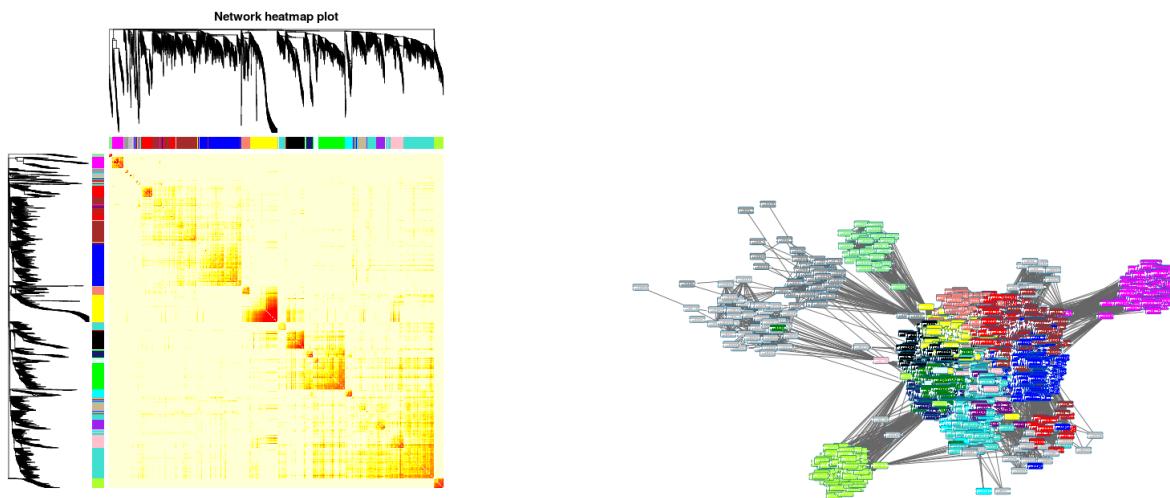
où a_{gs} est la mesure d'adjacence entre deux gènes g et s (équation 2.4), $l_{gs} = \sum_v a_{gv}a_{vs}$ est la connectivité des voisins communs entre g et s , et $k_g = \sum_v a_{gv}$ est la connectivité du gène g .

La mesure “topological overlap” permet ainsi de quantifier la co-expressivité entre deux gènes en prenant en compte la corrélation entre les expressions des deux gènes, mais aussi la corrélation entre l'expression des gènes qui, dans le réseau, sont reliés à la fois au premier et au deuxième gène.

La détection des modules s'effectue ensuite à partir de la TOM. Pour cela, WGCNA utilise la méthode de clustering hiérarchique ascendant sur la TOM. Le clustering hiérarchique ascendant est une méthode de classification automatique qui permet de rassembler des éléments en classes. La méthode rassemble les éléments qui sont proches, cette notion est définie par une mesure de distance. Dans le cas de WGCNA, les éléments sont les gènes, les classes sont les modules et la mesure de distance entre éléments est le topological overlap. On part d'un état où tous les gènes sont dans des modules différents, puis étape par étape on rassemble dans un même module les gènes qui, d'après la TOM, sont les plus proches. On obtient ainsi une hiérarchie, représentée graphiquement par un dendrogramme (comme sur les axes de la figure 2.7a).

WGCNA propose ensuite plusieurs méthodes pour décider d'un seuil de « coupe » de la hiérarchie qui va ainsi définir les modules. Lorsqu'on visualise le réseau à partir de la TOM avec une visualisation dirigée par la force, les modules peuvent s'observer : ce sont des groupes de gènes bien isolés les uns des autres (figure 2.7b).

Le logiciel WGCNA est accompagné d'un tutoriel, qui permet de voir comment utiliser les différentes



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

(b) Le réseau est représenté par Cytoscape grâce à une méthode dirigée par la force des arêtes du graphe.

FIGURE 2.7 – Réseau de gènes des données du tutoriel WGCNA.

fonctions. Le tutoriel fournit des données de puces à ADN avec lesquelles s'exercer. Afin de visualiser les modules, WGCNA propose un dendrogramme dans lequel les gènes sont organisés par le clustering hiérarchique et les modules sont représentés sous les gènes par différentes couleurs. WGCNA propose également une visualisation de la TOM grâce à une heat map avec les modules sur les axes. Les valeurs de la TOM des données du tutoriel peuvent être visualisées sur la figure 2.7, grâce à la heat map de WGCNA (figure 2.7a) et grâce au logiciel Cytoscape (figure 2.7b). La visualisation de WGCNA permet de voir quels sont les modules dans lesquels les gènes sont les plus corrélés. Une module intéressant est un module dans lequel les gènes du module sont corrélés entre eux, et faiblement avec les gènes des autres modules. On peut voir sur la figure 2.7a que le module jaune ainsi que le module vert qui se trouve à l'extrême droite du dendrogramme, sont intéressants puisque en face de ces modules, la heat map ne fait apparaître de fortes valeurs de la TOM (couleur rouge) qu'autour de la diagonale. La visualisation offerte par Cytoscape (figure 2.7b) a été réalisée en utilisant le « Force-directed layout » : plus le poids d'une arête est important, plus les nœuds sont proches. Dans cette visualisation, un gène est représenté par un nœud, et les arêtes sont étiquetées par la valeur de la TOM. Un seuillage sur les valeurs de la TOM a été fait afin de générer cette visualisation. Un ensemble de gènes intéressant dans cette visualisation est un groupe de gènes à l'écart des autres. On ne retrouve pas visuellement sur le réseau de la figure 2.7b les 22 modules détectés par WGCNA, puisque, comme on peut le voir sur la figure 2.7a, certains modules sont en réalité peu différenciés entre eux, et également puisque tous les liens ne sont pas présents à cause du seuil.

La complexité de WGCNA est en $\mathcal{O}(pn^2)$ pour le calcul de la corrélation, avec p le nombre d'échantillons et n le nombre de gènes. Le calcul de la TOM a une complexité de $\mathcal{O}(n^3)$.

En raison de cette complexité, WGCNA ne peut être exécuté sur un ensemble de gènes n très grand.

Si le nombre de gènes est trop important, WGCNA propose une construction du réseau par blocs. Avant la construction d'un réseau de gènes, WGCNA découpe les données en blocs grâce à la méthode des k -moyennes, k étant le nombre de blocs créés qui est déterminé en fonction de la taille des données. Les données sont l'expression des gènes dans plusieurs expériences, l'algorithme des k -moyennes va donc regrouper les gènes qui ont une intensité moyenne semblable tout au long des expériences. Une fois les blocs créés, WGCNA traite les blocs un à un, indépendamment, en calculant le réseau de gènes du bloc et détecte les modules à partir de ce réseau. On obtient alors autant de réseaux que de blocs, et les modules ne détectent que les gènes interconnectés d'un bloc, ce qui prive donc d'une partie de l'information contenue dans les données initiales.

2.3.2 Méthodes à base d'information mutuelle

L'information mutuelle [Cover et Thomas, 1991] est une mesure issue de la théorie de l'information et permet de mesurer la dépendance entre deux variables discrètes. L'information mutuelle entre deux variables aléatoires X et Y définies respectivement sur \mathcal{X} et \mathcal{Y} , avec une distribution de probabilités jointes $p(x, y)$ et des distributions marginales $p(x)$ et $p(y)$ est :

$$I(X, Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log\left(\frac{p(x, y)}{p(x)p(y)}\right) \quad (2.6)$$

Une propriété intéressante de l'information mutuelle est qu'elle est nulle si et seulement si les deux variables sont indépendantes. Mais l'estimation de l'information mutuelle à partir des données est difficile et plusieurs estimateurs ont été proposés pour cela [Steuer *et al.*, 2002]. Un estimateur qui prend en compte les valeurs continues a été proposé dans [Meyer *et al.*, 2007]. L'information mutuelle entre deux gènes g et s , où σ_g^2 et σ_s^2 sont les variances respectives de g et s , et $|C|$ est le déterminant de la matrice de covariance, est donc estimée par :

$$I(g, s) = \frac{1}{2} \log\left(\frac{\sigma_g^2 \sigma_s^2}{|C|}\right) \quad (2.7)$$

Plusieurs méthodes d'inférence sont basées sur l'information mutuelle (voir table 2.1); parmi elles, nous décrivons ici la méthode ARACNE, Algorithm for the Reconstruction of Gene Regulatory Networks [Margolin *et al.*, 2006], et la méthode C3NET, Conservative Causal Core Network [Altay et Emmert-Streib, 2010].

Dans ces deux méthodes d'inférence, il y a le même prétraitement sur les données permettant d'estimer l'information mutuelle au mieux : la copula transformation. La copula transformation [Steuer *et al.*, 2002, Kurt *et al.*, 2014] permet d'améliorer la performance des estimateurs en ne travaillant pas sur les valeurs d'expression elles-mêmes, mais sur leur rang. Après avoir remplacé chaque valeur par son rang dans le vecteur, on divise le rang par le nombre de valeurs dans le vecteur. C'est une normalisation qui permet ainsi de travailler sur des valeurs comprises entre 0 et 1.

Dans les deux méthodes, la première étape des algorithmes est de calculer l'information mutuelle

entre chaque paire de gènes.

Ensuite, une autre étape commune entre ARACNE et C3NET est le test de significativité de l'information mutuelle. Pour tester si une valeur d'information mutuelle est significative ou non, on fixe un seuil I_0 pour lequel on considère que toutes les valeurs inférieures sont non-significatives, et les valeurs supérieures sont significatives. Afin de déterminer ce seuil, on examine ce que serait la distribution des valeurs d'information mutuelle pour des données aléatoires, c'est-à-dire si les expressions des gènes ne sont pas issues d'un système de régulation. Les données d'expression sont donc permutées aléatoirement et la matrice d'information mutuelle associée est calculée. L'opération est répétée un certain nombre de fois, pour obtenir une distribution des valeurs d'information mutuelle dans le cas de données aléatoires. Le seuil I_0 est alors déterminé de telle sorte qu'un faible pourcentage (seuil de rejet α du test) des valeurs obtenues sur les données aléatoires soient supérieures à I_0 . Ainsi si $\alpha = 0.01$, le seuil I_0 est tel que 1% des valeurs d'information mutuelle issues des données aléatoires sont supérieures à I_0 .

Dans la méthode ARACNE, toutes les interactions dont l'information mutuelle est inférieure au seuil sont donc supprimées et on obtient donc la matrice d'adjacence pondérée du réseau.

La seconde étape d'ARACNE est un post-traitement visant à éliminer les interactions indirectes entre deux gènes qui résultent de l'interaction avec un troisième gène. Ce traitement se base sur la Data Processing Inequality (DPI). La DPI stipule que s'il existe une interaction entre deux gènes g_1 et g_3 qui s'effectue via un troisième gène g_2 , alors l'information mutuelle respecte l'inégalité suivante :

$$I(g_1, g_3) \leq \min[I(g_1, g_2); I(g_2, g_3)] \quad (2.8)$$

En utilisant cette inégalité, ARACNE évalue un à un chacun des triplets (g_1, g_2, g_3) et supprime l'interaction la plus faible. Ici, la notion de « triplet » fait référence à un graphe complet de taille 3, c'est-à-dire qu'il existe une interaction entre chaque paire de gènes du triplet. La figure 2.8 illustre cette étape d'ARACNE. En ne gardant que les valeurs significatives d'information mutuelle, on obtient un réseau de gènes (figure 2.8a). Dans ce réseau il existe deux triplets (g_1, g_2, g_3) et (g_2, g_3, g_4) , et dans chacun des triplets, de manière indépendante, l'interaction la plus faible est retirée (figure 2.8b et 2.8c). On obtient ainsi le réseau final (figure 2.8d).

On note que l'identification des triplets s'effectue avant toute suppression, ainsi, quelque soit l'ordre dans lequel sont traités les triplets, on obtiendra le même réseau final. Dans la figure 2.8, même si le lien entre g_2 et g_3 est supprimé dans l'analyse du triplet (g_1, g_2, g_3) , le triplet (g_2, g_3, g_4) est toujours considéré comme un triplet et le lien entre g_2 et g_4 est ainsi supprimé.

Pour un jeu de données constitué de n gènes et p échantillons, l'estimation de l'information mutuelle est en $\mathcal{O}(n^2 p^2)$ et l'utilisation de la DPI est en $\mathcal{O}(n^3)$. ARACNE a donc une complexité en $\mathcal{O}(n^3 + n^2 p^2)$.

C3NET est une autre méthode d'inférence de réseau permettant de construire un « cœur de réseau » dans lequel, pour chaque gène, seule la plus forte interaction est représentée. Les auteurs de C3NET sont partis de l'observation que les réseaux basés sur une mesure d'interaction paire à paire sont très dépen-

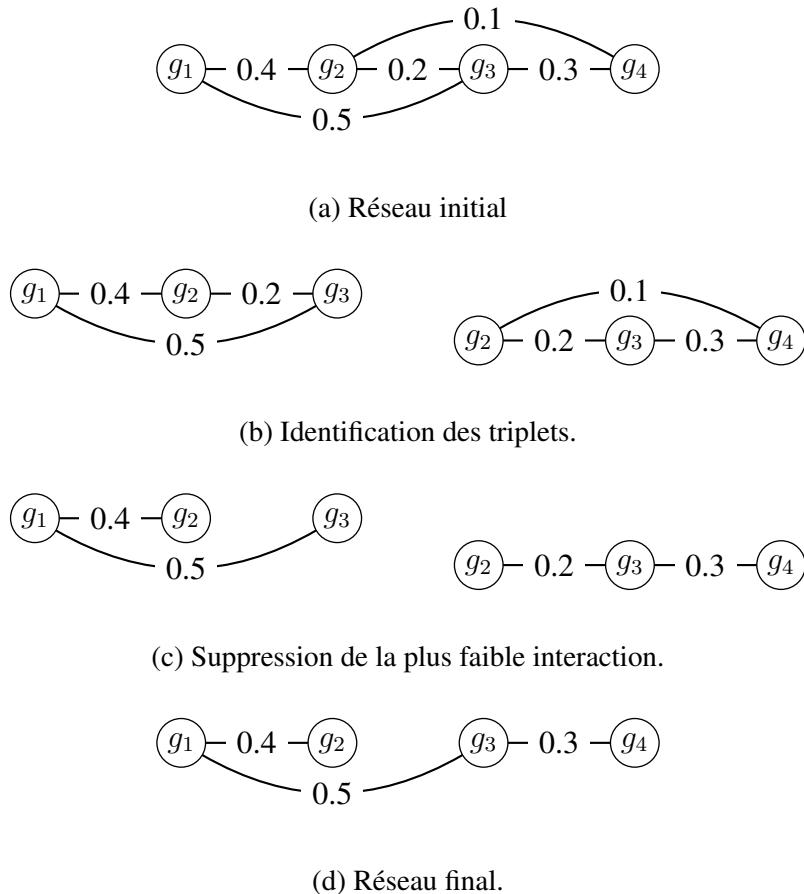


FIGURE 2.8 – Déroulement de la suppression des interactions indirectes dans ARACNE en utilisant la DPI.

dants des structures locales du réseau à reconstruire. Des structures dans le réseau, comme un simple lien ou des sous-réseaux connectés, sont inférées de manière très différente selon la méthode d'inférence utilisée. C'est pourquoi C3NET propose de ne représenter que le cœur de réseau, qui représente la structure principale du réseau.

Le fonctionnement de C3NET est illustré par la figure 2.9. La première étape de C3NET est de calculer l'information mutuelle entre toutes les paires de gènes. Ensuite, en ne gardant que l'information mutuelle statistiquement significative, C3NET construit une matrice de connectivité C . Lors de la deuxième étape, C3NET identifie l'information mutuelle maximale pour chacun des gènes. Cela fournit une matrice d'adjacence A_j d'un graphe orienté. Finalement, l'information mutuelle étant une mesure symétrique, C3NET transforme le réseau orienté en un réseau non-orienté A en transformant les arcs de A_j en arêtes.

Dans la figure 2.9, on remarque ainsi qu'à partir de la matrice d'information mutuelle I , on obtient la matrice de connectivité C où les valeurs d'information mutuelle non-statistiquement significatives sont associées à 0, toutes les autres valeurs sont considérées comme des interactions potentielles et associées à 1. Ensuite C3NET calcule l'interaction maximale pour chaque gène et on obtient alors la matrice d'ad-

$$I = \begin{pmatrix} 1.0 & 0.7 & 0.9 & 0.8 \\ 0.7 & 1.0 & 0.6 & 0.5 \\ 0.9 & 0.6 & 1.0 & 0.1 \\ 0.8 & 0.5 & 0.1 & 1.0 \end{pmatrix}$$

$$C = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix} \Rightarrow A_j = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \Rightarrow A = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

FIGURE 2.9 – Visualisation des principales étapes de C3NET.

jacience A_j . La matrice est représentée par un graphe orienté où un arc entre les nœuds 4 et 1 signifie que l’information mutuelle la plus élevée pour le gène 4 est celle avec le gène 1. Enfin le réseau est transformé en réseau non-orienté.

Le graphe C est le réseau de gènes obtenu avant l’application de C3NET, et le graphe A est le réseau final. La différence entre ces deux réseaux est l’arête entre 2 et 3 qui est présente avant l’application de C3NET mais pas après : cela signifie que l’information mutuelle entre les gènes 2 et 3 n’était pas la plus élevée ni pour le gène 2 ni pour le gène 3. En effet, le graphe A_j nous montre que l’information mutuelle maximale pour le gène 2, ainsi que pour le gène 3, est l’information mutuelle partagée avec le gène 1, d’où les arcs $2 \rightarrow 1$ et $3 \rightarrow 1$. L’interaction entre les gènes 2 et 3 ne fait donc pas partie du cœur de réseau et n’apparaît pas dans le réseau final de C3NET.

L’algorithme de C3NET consiste à parcourir une matrice $n \times n$, où n est le nombre de gènes, et de trouver le maximum pour chaque ligne. La complexité de cet algorithme est donc en $\mathcal{O}(n^2)$, auquel s’ajoute l’estimation de l’information mutuelle, ce qui donne une complexité de C3NET en $\mathcal{O}(n^2 + n^2 p^2)$. C3NET a une complexité assez faible mais crée des matrices carrées de taille conséquente.

2.3.3 Méthodes à base de régression

Plusieurs méthodes abordent le problème de la reconstruction d’un réseau entre n gènes, en considérant n problèmes de régression dans lesquels on cherche à expliquer le vecteur d’expression d’un gène cible à partir des vecteurs d’expression des autres gènes. Cette approche est particulièrement adaptée lorsqu’on cherche à étudier un réseau de transcription et que l’on connaît, parmi les gènes étudiés, lesquels sont des facteurs de transcription et lesquels sont des gènes cibles. Ainsi, dans des logiciels comme TIGRESS [Haury *et al.*, 2012] et NARROMI [Zhang *et al.*, 2013], on peut spécifier en entrée quel est l’ensemble des facteurs de transcription. Dans ces approches, différentes techniques, comme les méthodes de ré-

gression pénalisée, de sélection par stabilité, d'élimination récursive, sont également employées pour ne sélectionner dans le réseau que les interactions les plus significatives, et obtenir la topologie scale-free du réseau.

Nous détaillons ici la méthode GENIE3, GEne Network Inference with Ensemble of trees [Huynh-Thu *et al.*, 2010], qui résout chaque problème de régression à l'aide de forêts aléatoires.

La figure 2.10, tirée de [Huynh-Thu *et al.*, 2010], montre les différentes étapes d'exécution de GENIE3. GENIE3 transforme le problème d'inférer un réseau de gènes constitué de n gènes en n problèmes de régression dans lesquels l'algorithme va assigner un poids $w_{i,j}$ représentant l'influence du gène i dans l'expression du gène j . À partir des données d'expression, pour chaque gène j , un ensemble d'apprentissage est donc généré, et l'objectif est de construire un arbre de régression expliquant les valeurs du gène j .

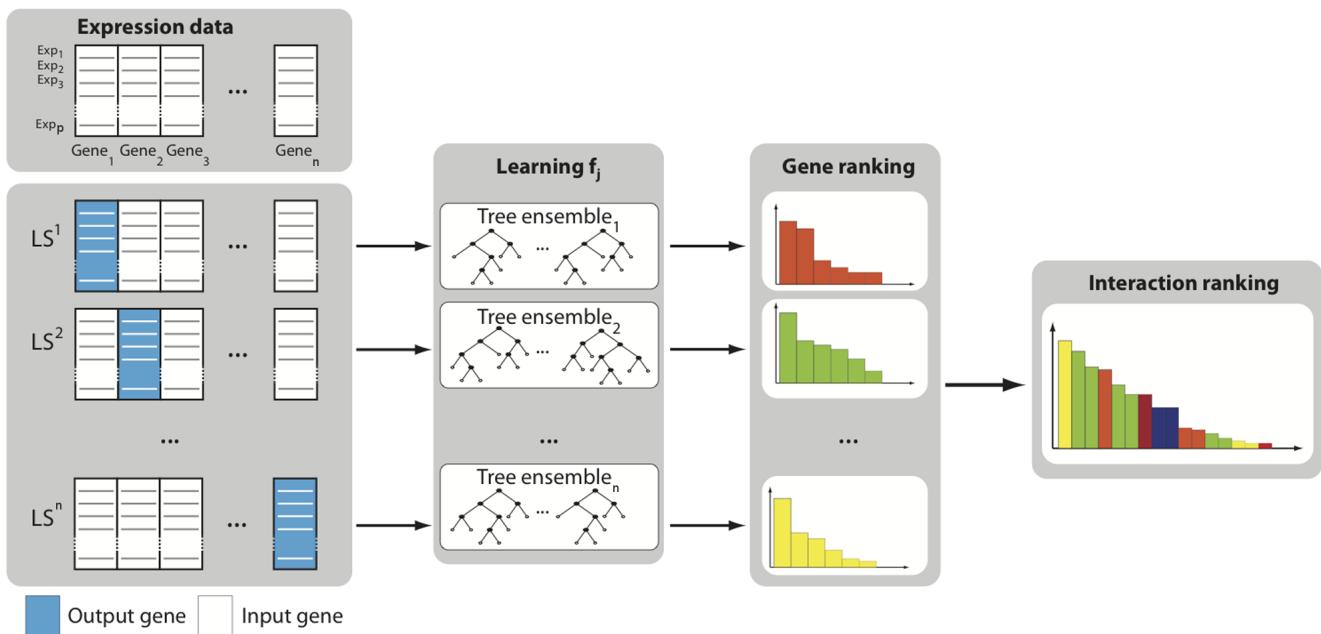


FIGURE 2.10 – Procédure de GENIE3. Pour chaque gène $j = 1, \dots, n$, un ensemble d'apprentissage LS^j est généré avec le niveau d'expression de j comme sortie et le niveau d'expression de tous les autres gènes comme entrées. LS^j permet d'apprendre une fonction f_j et un rang de tous les gènes sauf j est calculé. Les n rangs sont ensuite agrégés pour former un poids de tous les liens de régulation. Illustration issue de [Huynh-Thu *et al.*, 2010].

La méthode d'arbres de régression ne fournit qu'une solution dont on ne garantit pas l'optimalité. GENIE3 utilise donc la méthode des forêts aléatoires dans laquelle plusieurs arbres de régression sont construits de manière stochastique ("Learning f_j " dans la figure 2.10). La méthode des forêts aléatoires sélectionne aléatoirement K gènes parmi les gènes d'entrée pour réaliser un arbre de régression, et répète l'opération T fois. L'ensemble des poids $w_{i,j}$ ainsi obtenus pour un gène de sortie j est trié selon une mesure qui indique la pertinence de l'expression du gène i dans la prédiction de l'expression du gène j . Ce tri permet d'assigner un rang au gène i dans le problème de régression lié au gène j ("Gene ranking").

Finalement un poids de régulation est calculé en fonction des poids et des rangs issus de chaque ensemble d'apprentissage (“Interaction ranking” dans la figure 2.10). Le poids de l'interaction est orienté ($w_{i,j} \neq w_{j,i}$) et non signé, il représente la force de régulation du gène i sur le gène j ne précisant pas si i active ou inhibe j . Une interaction étiquetée $w_{i,j}$ est assignée pour tous les gènes d'entrée i utilisés par la méthode de la forêt aléatoire pour un gène de sortie j . L'utilisation d'un seuil sur la valeur d'interaction et/ou sur le rang de l'interaction permet enfin de construire le réseau.

La complexité de GENIE3 est en $\mathcal{O}(nTKp \log p)$, avec n le nombre de gènes, T le nombre d'arbres sélectionnés dans la méthode de forêt aléatoire, K le nombre d'attributs utilisés pour la méthode de régression et p la taille de l'échantillon d'apprentissage.

La régression est également utilisée dans la méthode h-LICORN [Chebil *et al.*, 2014], mais l'originalité de ce travail repose surtout dans l'algorithme LICORN [Elati *et al.*, 2007] sur lequel il repose. LICORN cherche à construire des réseaux de régulation coopératifs, dans lesquels plusieurs co-régulateurs agissent ensemble pour activer ou réprimer un gène cible. Le modèle de régulation coopérative utilisé dans LICORN est donné dans la figure 2.11.

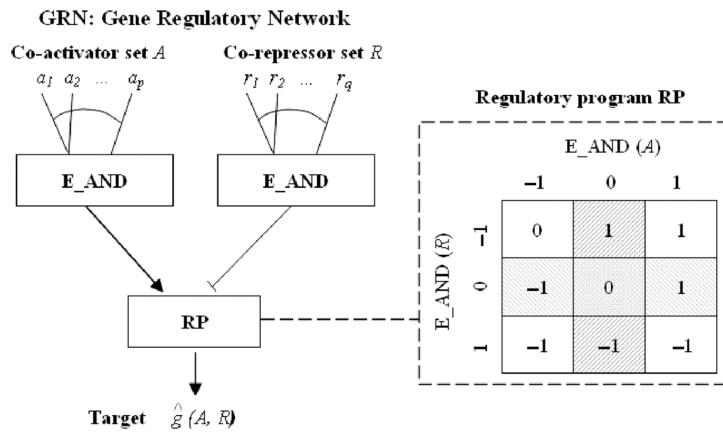


FIGURE 2.11 – Définition du programme de régulation RP utilisé par LICORN. Illustration issue de [Elati *et al.*, 2007].

LICORN travaille sur des données discrétisées dans l'ensemble $\{-1, 0, 1\}$ signifiant qu'un gène est respectivement sous-exprimé, sans changement, ou sur-exprimé. À partir d'un ensemble de gènes cibles et d'un ensemble de régulateurs (les facteurs de transcription), et de leurs matrices d'expression discrétisées, l'objectif est de trouver un réseau qui, pour chaque gène, explique le mieux ses valeurs d'expression en fonction de ce programme de régulation. Pour limiter l'espace de recherche à considérer pour déterminer les ensembles de co-activateurs et de co-répresseurs, LICORN utilise la méthode de recherche de motifs fréquents Apriori [Agrawal *et al.*, 1993], ou plus exactement une extension d'Apriori permettant de gérer les valeurs 1 mais aussi les valeurs -1 de la matrice d'expression. Pour chaque gène g , LICORN détermine alors des ensembles de régulateurs potentiels, en considérant les supports de ces ensembles fréquents et le supports de g . À partir de cet ensemble réduit de candidats potentiels, le meilleur réseau possible pour chaque gène est déterminé par une recherche exhaustive ; les différents réseaux sont es-

timées par une mesure de score qui compare, dans cet espace discret, les valeurs prédites et réelles de l'expression des gènes.

H-LICORN est une extension de LICORN qui travaille à la fois avec les données discrétisées et les données continues. En effet H-LICORN utilise une régression linéaire pour déterminer, parmi les réseaux candidats déterminés par LICORN, celui qui permet le meilleur ajustement des données. Les résultats obtenus montrent que la régression linéaire discrimine mieux les candidats que le score utilisé dans LICORN.

Puisque nous avons vu que LICORN utilise une méthode de fouille de données, nous pouvons signaler également la méthode RulNet [Vincent *et al.*, 2015] qui recherche des règles d'association représentant les interactions entre gènes. La méthode de recherche se fait en interaction avec l'utilisateur qui peut exprimer par des requêtes de type SQL les informations qui l'intéressent.

Nous avons présenté quelques méthodes de reconstruction de réseaux de gènes. Chacune de ces méthodes a été appliquée sur des données biologiques réelles pour répondre à une question biologique particulière. Mais auparavant chacune de ces méthodes a été évaluée sur des données synthétiques permettant un processus de validation. Nous décrivons dans le paragraphe suivant la manière dont cette validation peut être réalisée.

2.4 Générateurs de données

En utilisant des données réelles, il est impossible de comparer deux méthodes d'inférence puisqu'on ne connaît pas le « véritable » réseau biologique qu'elles doivent reconstruire. Pour évaluer la capacité des différentes méthodes d'inférence proposées dans la littérature, il faut se placer dans un contexte supervisé permettant de comparer un réseau inféré avec un réseau connu. Il faut donc pouvoir générer des données synthétiques à partir d'un réseau connu et d'un modèle proche du modèle biologique, et c'est ce que proposent les générateurs de données.

Le fonctionnement d'un générateur de données est schématisé dans la figure 2.12 : le principe est de créer un réseau *in silico* (**A**), de simuler une activité à ce réseau afin de générer des données d'expression (**B**), ces données servent pour les méthodes d'inférence (**C**). Grâce au réseau **A**, on peut évaluer les réseaux donnés par les méthodes d'inférence (**D**).

La première étape de la génération de données *in silico* s'effectue à partir d'un réseau biologique connu. Les générateurs utilisent un réseau déjà inféré et qui fait consensus dans la communauté biologique afin d'avoir une topologie la plus proche d'un vrai réseau biologique. Cela permet par la suite d'avoir des données d'expression les plus plausibles possible. Le réseau biologique connu permet donc de définir la structure du réseau final.

La seconde étape de la génération de données *in silico* permet de simuler une activité aux gènes

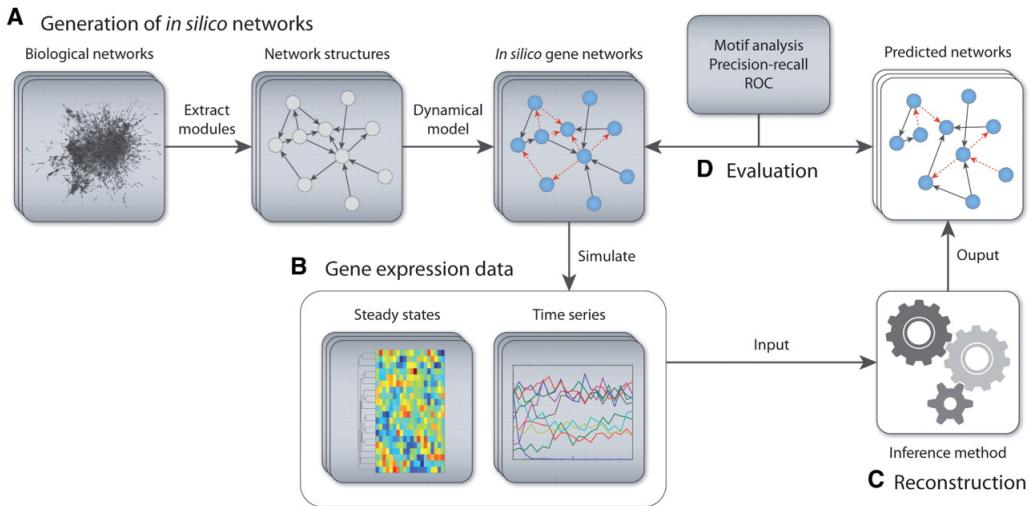


FIGURE 2.12 – Organisation de la génération des données simulées par GeneNetWeaver. Illustration issue de [Schaffter *et al.*, 2011].

du réseau. Un modèle mathématique associé au réseau va permettre de simuler ces données. Il existe deux types de données d'expression : les données à état stationnaire (steady-state) et les données de série temporelle (time series). Les données steady-state sont prélevées sur un organe à un moment dans une condition précise. Ce type de données est utilisé généralement pour décrire une condition : un individu malade, un individu sain, *etc*. On peut ainsi comparer deux individus distincts dans la condition. Les données time series sont des données qui décrivent une dynamique : c'est une succession de données d'expression à un intervalle de temps donné. Ce type de données est utilisé pour étudier l'évolution d'une processus, ainsi on compare l'individu avec lui-même mais dans un état différent. Cette seconde étape fournit donc des données d'expressions simulées à partir du réseau de référence. Ces données d'expressions sont utilisées comme données d'entrée pour les méthodes d'inférence à évaluer.

Il existe plusieurs générateurs de données biologiques. Nous présentons ici deux générateurs étudiés dans la thèse : SynTReN, que nous utilisons pour comparer les performances de la méthode que nous avons développée, et GNW qui a été utilisé afin de produire les données pour le challenge « In Silico Network » des compétitions internationales DREAM3 et DREAM4 (Dialogue for Reverse Engineering Assessments and Methods). Le DREAM4 a d'ailleurs été remporté par la méthode GENIE3.

Synthetic Transcriptional Regulatory Networks (SynTReN) [Bulcke *et al.*, 2006] est un générateur de données steady-state *in silico*. La génération de données avec SynTReN s'effectue en trois étapes. La première étape est l'extraction d'un sous-réseau à partir d'un réseau biologique connu ; la seconde étape est la sélection d'un réseau de fond afin de brouter les données ; la troisième étape est la simulation des données.

SynTReN propose de sélectionner un sous-réseau à partir des réseaux de deux organismes : la bactérie *Escherichia coli* ou la levure *Saccharomyces cerevisiae*. La sélection d'un sous-réseau peut s'effectuer à l'aide de deux méthodes différentes : la méthode par sélection du voisin (*neighbour addition*) ou la mé-

thode par les clusters (*cluster addition*). La méthode par sélection du voisin choisit un nœud aléatoirement dans le réseau de départ, puis un voisin des nœuds du sous-réseau est sélectionné aléatoirement pour être ajouté à son tour dans le sous-réseau. On obtient ainsi un sous-réseau connexe du réseau d'origine. La méthode par les clusters choisit également un nœud aléatoirement dans le réseau de départ, mais il ajoute ce nœud ainsi que l'ensemble de ses voisins dans le sous-réseau. Par la suite un autre nœud, voisin d'un des nœuds du sous-réseau, est incorporé avec l'ensemble de ses voisins dans le sous-réseau. On obtient là aussi un sous-réseau connexe du réseau d'origine. La différence entre ces deux méthodes de sélection est qu'un sous-réseau obtenu par la méthode *cluster addition* sera plus dense : les nœuds sélectionnés seront plus inter-connectés que dans un sous-réseau obtenu par la méthode *neighbour addition*.

La particularité de SynTReN est le réseau de fond, c'est un réseau qui ne possède aucune connexion avec le sous-réseau extrait, les gènes qui le composent vont s'exprimer afin de brouter les données générées. Dans une étude transcriptomique, il est admis que l'ensemble du génome n'est pas impliqué dans la condition étudiée [Yang *et al.*, 2002]. Les gènes du réseau de fond permettent ainsi de modéliser des mécanismes biologiques qui ne sont pas influencés par la condition expérimentale. L'expression des gènes de fond est fournie par SynTReN, il faut donc que la méthode d'inférence puisse identifier non seulement les régulations qui s'effectuent entre les gènes du réseau *in silico*, mais il faut également qu'elle puisse différencier les gènes qui appartiennent à ce réseau des gènes de fond.

GeneNetWeaver (GNW) [Schaffter *et al.*, 2011] est un générateur de données steady-state ou time series *in silico*. Le fonctionnement de GNW est schématisé par la figure 2.12. GNW propose également d'extraire un sous-réseau à partir des réseaux d'*E. coli* ou de *S. cerevisiae*. La méthode utilisée pour sélectionner le sous-réseau est la méthode d'extraction de modules [Marbach *et al.*, 2009]. L'extraction de module choisit un nœud aléatoirement, puis ajoute tous les voisins des gènes contenus dans le sous-réseau qui ont la plus forte modularité Q . La modularité est calculée comme étant le quotient entre le nombre d'interactions dans le sous-réseau par le nombre de ces interactions présentes dans un réseau aléatoire. GNW permet non seulement la génération de données à état constant ou des séries temporelles, mais il permet également de simuler différentes opérations biologiques telles que le knock-out ou le knock-down. Ces opérations sont utilisées en biologie afin de pouvoir mettre en évidence le rôle de régulation d'un gène. Ainsi GNW permet de simuler aisément une condition expérimentale spécifique.

2.5 Analyse différentielle de réseaux

Après avoir rappelé les notions essentielles sur les réseaux de gènes et présenté la problématique et les méthodes d'inférence de réseaux, nous discutons maintenant de travaux qui s'appuient sur une méthodologie d'analyse différentielle de réseaux, c'est à dire de comparaison de réseaux.

Pour traiter les données à large échelle fournies par les méthodes modernes d'analyse du vivant (puces à ADN, ChIP-on-chip, RNA-Seq, *etc.*), l'analyse comparative des données observées dans des conditions

expérimentales différentes est un élément clé, elle permet l'identification d'une partie des acteurs impliqués dans le processus biologique étudié.

Ainsi, la visualisation offerte par les puces en hybridation comparative et l'analyse de l'expression différentielle des gènes permettent l'identification des gènes ayant une activité significativement différente d'une condition à l'autre. En médecine, la confrontation de données obtenues sur des patients sains d'une part et sur des patients malades d'autre part est la base de nombreuses études. Dans [Golub *et al.*, 1999], les auteurs ont montré que la reconnaissance de différents types de leucémie pouvait être faite à partir d'un petit ensemble de gènes ayant des niveaux d'expression très différents au sein des échantillons de moelle épinière ou de sang prélevés. En s'appuyant sur des techniques d'apprentissage automatique, ces approches sont utiles pour le diagnostic ou le pronostic de la maladie, mais elles ne fournissent pas d'explications sur les dysfonctionnements causant la maladie.

C'est pourquoi des travaux récents, dont une revue est disponible dans [de la Fuente, 2010], s'intéressent à l'analyse différentielle de réseaux afin d'identifier les régulations au sein de la cellule qui sont altérées dans les échantillons malades.

Dans [Sharan et Ideker, 2006, Ideker et Krogan, 2012], la méthodologie de comparaison de réseaux biologiques est présentée suivant trois modes d'analyse possibles : l'alignement de réseaux qui concerne au moins deux réseaux de même type dans des espèces différentes, l'intégration de réseaux qui met en jeu des réseaux de différents types pour la même espèce et l'interrogation à partir de sous-réseaux connus. Ils prédisent que, comme ce fut le cas pour le traitement des séquences, la comparaison de réseaux va dépendre d'avancées méthodologiques et techniques, telles que des algorithmes de recherche performants, des techniques d'alignements multiples, des intégrations efficaces des bases de données publiques, *etc.*

L'alignement de réseaux d'espèces différentes est assez éloigné des problèmes que nous avons voulu traiter. Nous citerons seulement un exemple caractéristique de ces travaux, PathBlast [Kelley *et al.*, 2003], qui permet l'alignement de deux réseaux protéines-protéines en tenant compte de la topologie des réseaux (interactions communes) mais aussi de la similarité entre les séquences des protéines en interaction. Comme dans le cas de l'alignement de séquences, l'algorithme d'alignement de graphes permet la prise en compte de "gaps" et de "mismatches" et s'appuie sur des méthodes de programmation dynamique pour affecter un score aux différents alignements possibles.

Pour les autres situations, l'analyse différentielle des réseaux biologiques [Sharan et Ideker, 2006, Gill *et al.*, 2014] fait référence à l'ensemble des travaux qui étudient les changements qui peuvent être observés dans les réseaux d'interactions représentant un système biologique dans différentes conditions environnementales, différents types de tissus, différents états de maladie, *etc.*

Passer de l'analyse différentielle de l'expression à l'analyse différentielle de réseaux peut se faire

selon plusieurs approches.

Un premier type d'approche dans l'analyse différentielle de réseaux biologiques est d'intégrer des connaissances sur des réseaux d'interactions validés avec des données mesurant l'expression des acteurs de ces réseaux dans des situations biologiques particulières. Ainsi dans [Ideker *et al.*, 2002], l'approche proposée consiste à analyser un réseau connu d'interactions moléculaires afin d'identifier des sous-réseaux actifs dans des conditions particulières ; un sous-réseau actif est une région connexe du réseau qui présente des niveaux d'expression significativement différents dans les conditions comparées. Après avoir défini une mesure du score d'activité d'un sous-réseau, la recherche des sous-réseaux de plus forte activité est réalisée par une méthode de recuit simulé. Pour des réseaux protéine-protéine ou protéine-ADN de la levure, cette étude identifie des mécanismes de régulation pertinents. Contrairement aux méthodes de clustering utilisées habituellement pour traiter la matrice d'expression des gènes, l'intégration de la connaissance sur les interactions moléculaires permet de faire intervenir dans le sous-réseau des gènes qui n'ont pas de changement significatif, mais qui relient des gènes différemment exprimés.

Pour mettre en œuvre ces approches, on a donc besoin de scores et d'algorithmes appropriés pour identifier des sous-réseaux pertinents.

Il est également possible de concentrer l'étude sur les différences topologiques observées dans les réseaux. Ainsi, dans [Faisal et Milenković, 2014], la mise en relation d'un réseau d'interactions protéine-protéine avec des données d'expression liées à l'âge est utilisée pour construire des réseaux spécifiques à une tranche d'âge. Une interaction du réseau de référence est mise dans le réseau associé à un âge si les deux protéines impliquées sont exprimées dans les données des patients de cet âge. L'originalité de ce travail est de comparer les 37 réseaux d'âges différents ainsi obtenus suivant leurs propriétés topologiques. L'analyse globale de la topologie des réseaux ne montre pas de changements particuliers avec l'âge, mais des mesures plus locales, telles que le degré des nœuds ou le coefficient de clustering, révèlent qu'un ensemble de protéines voit ses modifications d'interactions corrélées avec l'âge. L'accumulation des connaissances sur les interactions d'une part, et des données disponibles d'autre part, ouvre la possibilité de nombreuses approches exploratoires de ce type.

Une autre approche possible est d'utiliser les données de différentes conditions pour inférer plusieurs réseaux qui sont alors comparés pour identifier les *gènes différemment connectés*. Ces différents travaux se distinguent par la manière dont ils sélectionnent les gènes étudiés, la manière de mesurer la corrélation et d'inférer les réseaux, la manière de considérer et de calculer les différences entre les réseaux.

Dans [Odibat et Reddy, 2012], des mesures de connectivité et de centralité et un algorithme inspiré de PageRank sont utilisés pour comparer des réseaux de corrélation calculés grâce à l'information mutuelle. Pour identifier les réponses cellulaires à différentes situations, des études récentes proposent d'identi-

fier les différences de co-expression en comparant plusieurs réseaux [Barabási *et al.*, 2011]. La comparaison peut identifier des sous-réseaux contenant des modifications de régulation significatives ; dans [Langfelder *et al.*, 2011], les auteurs mesurent la préservation des modules de réseaux inférés à partir de différentes conditions.

Les auteurs de C3NET proposent une méthode d'analyse différentielle de réseaux : Differential C3NET (DC3NET) [Altay *et al.*, 2011]. La méthode compare deux cœurs de réseaux obtenus grâce à la méthode C3NET : un réseau contrôle (C) et un réseau traitement (T). L'analyse différentielle produit trois réseaux différents : le réseau commun, le réseau différentiel contrôle et le réseau différentiel traitement.

Le réseau commun contiendra toutes les interactions qui sont significativement communes aux deux réseaux C et T. Le réseau différentiel contrôle ne contiendra que les interactions du réseau C qui ne sont pas significativement présentes dans le réseau T. De manière symétrique, le réseau différentiel traitement ne contiendra que les interactions du réseau T qui ne sont pas significativement présentes dans le réseau C.

Les réseaux C et T ont été inférés à partir de C3NET, il n'y a donc pour chaque gène qu'une seule interaction d'inférée. Pour l'analyse différentielle, DC3NET ne souhaite pas simplement vérifier la présence ou l'absence d'une interaction dans l'un et l'autre des réseaux. Afin de définir si une interaction est significativement absente, DC3NET évalue une interaction sur deux critères : la valeur d'information mutuelle et le rang de l'interaction. Ainsi, à l'aide de seuils, DC3NET considère que l'interaction entre a et b est présente si la valeur d'information mutuelle $I_{a,b}$ est parmi les meilleures de l'ensemble du réseau ou si l'interaction est parmi les meilleures interactions de a ou de b .

Dans [Altay *et al.*, 2011], cette méthode d'analyse différentielle est appliquée sur des données de cancer de la prostate, contenant 52 échantillons de tumeurs et 50 échantillons de tissus normaux pour 12 600 gènes analysés. L'inférence sur les échantillons de tumeur a fourni un réseau de 9 653 interactions, alors que l'inférence sur les données de tissus sains a donné un réseau de 8 930 interactions. L'analyse différentielle a conduit à un réseau commun contenant 992 interactions, un réseau TumeurDiff (interactions propres aux échantillons malades) de 2 409 interactions et un réseau NormalDiff de 2 025 interactions. Peu de relations de ces différents réseaux, 54 au total, sont retrouvées dans les bases d'interactions physiques telles que HPRD (Human Protein Reference Database) ou BIOGrid. Cela témoigne du fait que les relations physiques entre protéines sont difficiles à inférer à partir des données transcriptomiques. Une analyse bibliographique de certaines des interactions spécifiques à TumorNet identifie des associations bien connues pour jouer un rôle dans les phases du développement cellulaire et dont le rôle dans le développement tumoral mérite d'être exploré. Beaucoup de travaux proposent d'identifier des bio-marqueurs de maladie ou de la progression d'une maladie par une analyse différentielle de réseaux de corrélation en détectant des corrélations différentes dans des réseaux [Bockmayr *et al.*, 2013].

Dans le même ordre d'idée, l'outil Diffany [Landeghem *et al.*, 2016] propose à partir d'un réseau de

référence la construction et la visualisation d'un ensemble de réseaux spécifiques à des conditions expérimentales. Pour traiter différents types de réseaux, cette méthode intègre une ontologie sur les types d'interactions : interactions physiques, interactions dirigées ou non, pondérées ou non, *etc.* Cet outil propose la construction d'un réseau consensus commun aux différentes conditions et des réseaux différentiels qui mettent en évidence la réorganisation des liens (ou « recâblage ») propre à chaque condition.

Une autre approche est proposée par les auteurs de WGCNA [Langfelder et Horvath, 2008]. Pour rappel, WGCNA infère un réseau de gènes puis analyse la topologie du réseau en regroupant les gènes dans des modules. Pour comparer les réseaux, WGCNA va ainsi comparer les modules. Une première méthode propose de regarder la correspondance entre les modules : savoir combien de gènes les modules des deux réseaux ont en commun. On peut alors visualiser si le module d'un réseau correspond exactement à un module de l'autre réseau, ou si les gènes du module d'un réseau se retrouvent dans plusieurs modules de l'autre réseau. Pour mesurer la préservation d'un module, WGCNA utilise la mesure Z summary [Langfelder *et al.*, 2011] qui prend en compte la taille du réseau, la taille du module et le nombre d'échantillons entre autres. Cette préservation est indiquée en plus de la correspondance des gènes. Une seconde méthode propose de construire des modules consensus. Un module consensus est constitué de gènes qui sont dans un même module dans les deux réseaux étudiés. Le module consensus représente ainsi les points communs entre les différents réseaux. Pour obtenir les particularités de chacun des réseaux, on peut réaliser une correspondance entre les modules d'un réseau avec les modules consensus.

Dans toutes les approches décrites à ce jour, les méthodes d'analyse différentielle de réseaux comparent des réseaux impliquant les mêmes acteurs dans différentes conditions. Dans le cadre de cette thèse, nous souhaitons prendre en compte les transcrits anti-sens et étudier leur impact au sein d'une même condition. Nous allons donc travailler avec une condition mais deux ensembles d'acteurs. Nous nous inspirons donc des méthodes d'analyses différentielles de réseaux existantes afin de développer une nouvelle méthode capable de gérer des ensembles d'acteurs différents et de comparer les réseaux ainsi obtenus .



3

Données d'expression du Pommier

Dans toute analyse en bio-informatique, il est important de connaître l'origine et le contexte expérimental dans lequel les données ont été obtenues. La récolte et le stockage de ces informations appelées métadonnées sont cruciaux pour l'interprétation des résultats. Dans ce chapitre nous présentons les données que nous avons utilisées au cours de cette thèse et qui sont issues d'expériences menées sur le pommier, à l'Institut de Recherche en Horticulture et Semences.

Nous donnons tout d'abord quelques informations biologiques essentielles sur le pommier et sur les méthodes d'acquisition des données d'expression du pommier. Ensuite nous présentons les motivations qui nous ont conduits à utiliser ces données pour mener une étude sur le rôle des transcrits anti-sens. Enfin nous complétons ce chapitre par des statistiques descriptives de nos données dans un premier temps, puis par une analyse de type fouille de données dans un second temps.

3.1	Pommier et puce AryANE	60
3.2	Motivations	60
3.3	Définition des données	63
3.4	Statistiques descriptives	67
3.5	Première étude de réseau sur les données	75

3.1 Pommier et puce AryANE

Le pommier, de son nom latin *Malus x domestica*, est l’arbre fruitier le plus cultivé en France. Il appartient à la famille des *Rosaceae* qui regroupe des plantes herbacées comme le rosier, ainsi que des arbres comme le prunier ou le pommier. Le pommier est un organisme classiquement diploïde composé de 17 chromosomes. Son génome a été séquencé pour la première fois en 2010 [Velasco *et al.*, 2010], sa taille est estimée à 750 Mb et 57 386 gènes putatifs ont été identifiés. Les gènes du pommier sont classiquement nommés « MDPxxx » où « xxx » représente un numéro de 10 chiffres permettant de l’identifier.

L’Institut de Recherche en Horticulture et Semences a développé, dans le cadre du projet AryANE, une puce à ADN à partir du génome du pommier : la puce AryANE v1.0 [Celton *et al.*, 2014]. Cette puce couvre l’ensemble des transcrits codants prédicts, et 153 précurseurs de micro-ARN. La puce est réalisée par la société Nimblegen et est constituée de 135 000 sondes. Ces dernières sont des oligonucléotides de 60-mers, ce qui signifie que la séquence d’ADN simple brin est constituée de 60 nucléotides. Parmi les sondes de la puce, 126 022 sondes sont utilisées dans notre étude : 63 011 sondes sont associées à un sens prédict, et la puce a la particularité de posséder les 63 011 sondes complémentaires permettant de détecter l’expression du transcript complémentaire (anti-sens).

La présence de sondes mesurant l’expression des transcrits anti-sens est une spécificité de la puce AryANE puisque peu de puces à ADN ont été conçues pour analyser ces transcrits. Cela rend difficile l’acquisition de données publiques contenant le niveau d’expression des transcrits anti-sens pour d’autres organismes.

Chaque sonde possède une note de spécificité de 1 à 4, en fonction de la spécificité décroissante de la sonde : 100% spécifique, 20-mers conservés, 40-mers conservés ou 60-mers conservés, permettant une cross-hybridation plus ou moins importante, c’est-à-dire que des transcrits d’un autre gène ou d’un autre précurseur de micro-ARN viennent s’hybrider sur la sonde. Cette note de spécificité représente la confiance dans le fait que les transcrits mesurés soient bien ceux ciblés. Une sonde d’une spécificité 4 indique donc que l’ensemble de la sonde peut être hybridée par un autre transcript que celui ciblé. Les sondes anti-sens sont complémentaires des sondes sens et ont donc la même spécificité.

Notons qu’un nouveau séquençage du génome du pommier a été finalisé à l’IRHS en juillet 2017 [Daccord *et al.*, 2017] ; il estime à 651 Mb la taille du génome et identifie 44 105 gènes putatifs. Lors de cette thèse nous avons travaillé à partir de la première version du génome, c’est-à-dire celle de 2010. Nous discuterons dans la partie perspectives comment les traitements que nous avons réalisés pourraient être repris en tenant compte du nouveau génome.

3.2 Motivations

Grâce à la puce AryANE, l’IRHS a conduit une étude [Celton *et al.*, 2014] en utilisant des données d’expression sens et anti-sens sur différents organes du pommier : la fleur, le fruit 100 jours après anthèse¹,

¹L’anthèse est la période pendant laquelle la fleur est complètement ouverte et fonctionnelle.

le fruit à la récolte, la feuille, la racine, la tige, la plantule et les graines. Les données ont été extraites à partir de deux répétitions biologiques. Cette étude a été menée pour étudier la régulation des gènes qui influent sur le phénotype de la plante *via* l'étude du transcriptome. Les transcrits anti-sens ont été intégrés dans l'étude pour savoir s'ils jouent le même rôle dans le pommier, plante pérenne, que dans des plantes annuelles comme le maïs ou Arabidopsis [Vanhée-Brossollet et Vaquero, 1998, Ma *et al.*, 2006].

Pour réduire d'éventuels biais techniques dus à la présence de sondes non-spécifiques (sondes de spécificité 4), l'étude de [Celton *et al.*, 2014] a été menée uniquement sur les sondes dont la spécificité est comprise entre 1 et 3, soit un ensemble de 96 120 sondes. La production d'un tel atlas d'expressions a permis de mesurer pour l'ensemble du génome, la quantité de transcrits sens et de transcrits anti-sens exprimés dans au moins une des conditions. Le premier résultat marquant de cette étude est que 65% des gènes voient leur transcrit anti-sens exprimé lorsque leur transcrit sens l'est également. Cela est un résultat assez inattendu car dans d'autres organismes le pourcentage est plus faible ; ainsi chez Arabidopsis une étude [Wang *et al.*, 2005] a observé 30% des transcrits anti-sens exprimés lorsque le transcrit sens l'est aussi.

Pour avoir une caractérisation du transcriptome du pommier, l'étude a recensé le pourcentage d'expression de transcrits sens et de transcrits anti-sens selon le rôle fonctionnel putatif des gènes dans la cellule. La figure 3.1 montre le résultat de cette étude. Sur l'ensemble des sondes, pour les huit organes, il y a plus de transcrits sens exprimés que de transcrits anti-sens. Les gènes qui sont impliqués dans la traduction sont ceux qui ont le plus de transcrits sens et anti-sens exprimés. On remarque que pour certains organes (graine, plantule et les fruits) il y a plus de transcrits anti-sens exprimés que de transcrits sens. Enfin, les gènes qui sont impliqués dans la défense de la cellule ont, pour les huit organes, plus de transcrits sens exprimés que de transcrits anti-sens, mais lorsqu'on regarde le niveau d'expression moyen des transcrits, on observe que le niveau d'expression de l'anti-sens est très proche de celui du sens par rapport à l'ensemble des gènes. Cette dernière catégorie de gènes est la seule qui montre ce comportement. Dans cette étude il est ainsi montré que les transcrits anti-sens sont plus exprimés pour certaines fonctions biologiques que pour d'autres.

En plus de cette distinction selon le rôle fonctionnel putatif des gènes, l'étude a regardé l'expression des transcrits selon les organes dans lesquels ils sont exprimés. Cette analyse cherche à déterminer si les transcrits anti-sens sont spécifiques à certains organes, ou s'ils sont ubiquitaires c'est-à-dire qu'ils jouent un rôle dans plusieurs types de cellules différents. Les fonctions ubiquitaires sont généralement liées au fonctionnement générique d'une cellule, comme la respiration cellulaire. La figure 3.2 montre le ratio d'expression sens/anti-sens dans les différents organes étudiés. Il ressort de cette analyse qu'il y a plus de transcrits anti-sens que de transcrits sens pour les graines, la plantule et les plantes (100 jours après anthèse et à la récolte). À l'inverse, les organes comme les racines, la tige, les feuilles et les fleurs ne semblent pas avoir beaucoup de transcrits anti-sens exprimés par rapport aux transcrits sens. De plus, on remarque que le ratio sens/anti-sens penche très largement du côté des sens lorsqu'il s'agit de gènes qui sont exprimés dans les huit organes. Il semble donc que la transcription anti-sens soit spécifique à certaines organes, et que le niveau d'expression anti-sens soit plus important pour des gènes spécifiques

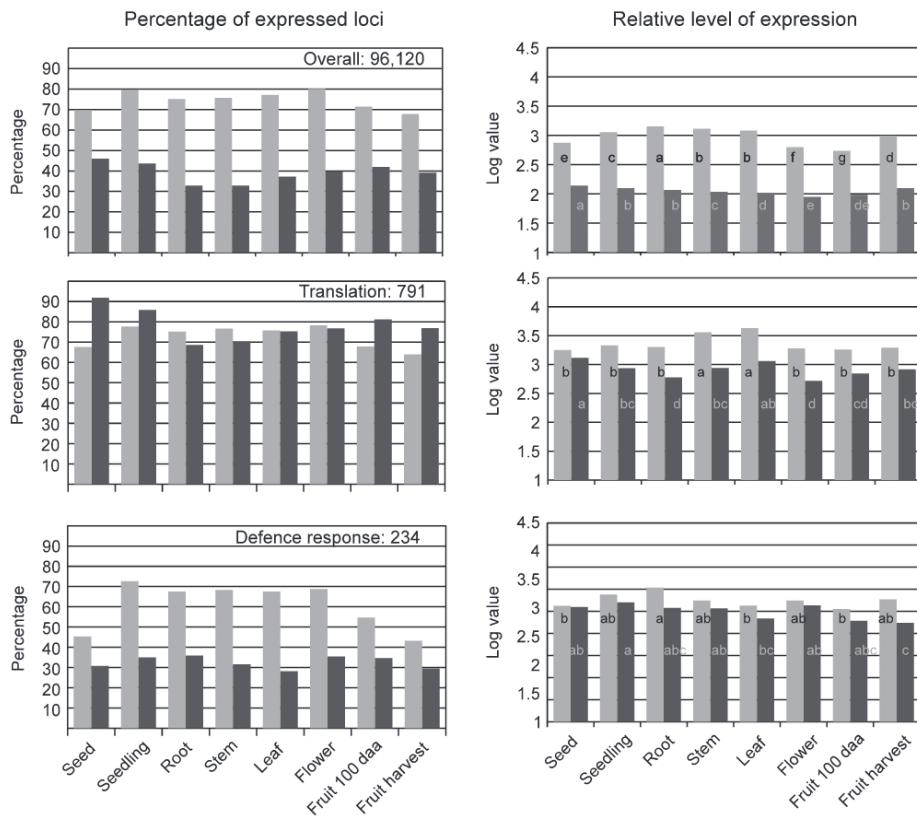


FIGURE 3.1 – Pourcentages de loci exprimés (gauche) et niveaux d'expression moyens (droite). Le premier graphique concerne l'ensemble des couples de loci de la puce AryANE. Le second graphique concerne les couples de loci reliés à des gènes ayant un rapport avec la traduction, ce qui représente 791 loci. Le troisième graphique concerne les couples de loci reliés à des gènes ayant un rapport avec la défense, ce qui représente 234 loci. Les transcrits sens sont en gris clair et les transcrits anti-sens en gris foncé. “daa” signifie “jours après anthèse”. Graphiques issus de [Celton *et al.*, 2014].

que pour des gènes impliqués dans des fonctions physiologiques ubiquitaires.

Enfin, plusieurs études [Borsani *et al.*, 2005, Wang *et al.*, 2005] indiquent qu'un lien existe entre la présence de transcrits anti-sens et celle des petits-ARN, qui interviennent dans le contrôle post-transcriptionnel. Afin de vérifier le rôle des transcrits anti-sens chez le pommier, les auteurs de [Celton *et al.*, 2014] ont séquencé les petits-ARN pour les échantillons du fruit à la récolte. Les auteurs ont ensuite regardé les petits-ARN qui correspondaient aux transcrits anti-sens, c'est-à-dire les transcrits anti-sens dont la séquence contient la séquence du petit-ARN. L'étude a été menée sur plusieurs tailles de petits-ARN, variant de 20 à 44 nucléotides. La figure 3.3 montre le pourcentage de transcrits correspondants avec des petits-ARN de 21 à 23 nucléotides. Sur ce graphique on observe que le pourcentage de transcrits qui correspondent avec un petit-ARN est plus important pour les couples tels que le transcrit sens et le transcrit anti-sens ont un niveau d'expression élevé. Lorsque le transcrit sens est exprimé mais que le transcrit anti-sens ne l'est pas, on observe peu de correspondance entre les transcrits et les petits-ARN. De même lorsque le transcrit anti-sens est exprimé mais pas le transcrit sens. C'est donc la présence conjointe à la fois du transcrit sens et du transcrit anti-sens qui voit le plus de correspondance

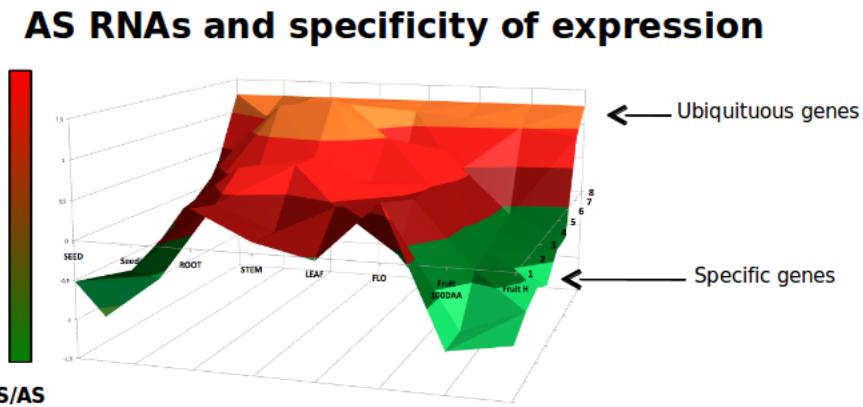


FIGURE 3.2 – Ratio d’expression sens/anti-sens dans les différents organes du pommier en fonction de la spécificité d’expression des gènes (1 : expression dans un seul organe ; 8 : expression dans les huit organes). Si le ratio est dans le vert, cela signifie que l’anti-sens est plus exprimé que le sens. À l’inverse, un ratio rouge indique une plus forte expression du sens que de l’anti-sens. Les huit organes sont placés dans l’ordre suivant : graine, plantule, racine, tige, feuille, fleur, fruit 100 jours après anthèse, fruit à la récolte.

entre les transcrits et les petits-ARN. L’étude montre ainsi qu’il existe une corrélation entre la présence de transcrits anti-sens et de petits-ARN (de 21 à 23 nucléotides et plus de 30 nucléotides).

Les différentes observations résumées ci-dessus ont motivé l’étude que j’ai menée dans ma thèse. En effet, les ARN anti-sens sont des ARN non-codants qui sont donc souvent négligés dans les études. Peu de travaux mesurent d’ailleurs leur transcription [Wang *et al.*, 2005, Maruyama *et al.*, 2012, Celton *et al.*, 2014]. Il est ainsi intéressant de faire une étude exploratoire de leur impact lorsqu’on les intègre dans des études de transcriptomique et notamment dans l’inférence de réseaux de gènes.

3.3 Définition des données

Nous avons ainsi décidé de travailler sur des données pommier concernant l’organe « fruit » puisque l’étude que nous venons de rappeler a montré que cet organe est très concerné par la transcription anti-sens.

Concernant les conditions expérimentales, nous avions à notre disposition les données de différentes expériences réalisées au sein de l’IRHS. L’institut s’intéresse notamment à la texture du fruit, un élément important pour les consommateurs, et il s’agit par exemple d’étudier comment la maturation du fruit peut agir sur ses qualités gustatives. L’atlas s’est donc étoffé en données sur le fruit à plusieurs stades de maturation. La maturation désigne le processus d’évolution du fruit ; après sa récolte, de plus le fruit est stocké dans des chambres froides ce qui constitue une situation de stress de froid auquel les cellules doivent s’adapter.

Nous avons donc considéré que ce contexte expérimental était pertinent pour aborder la problématique de la transcription anti-sens.

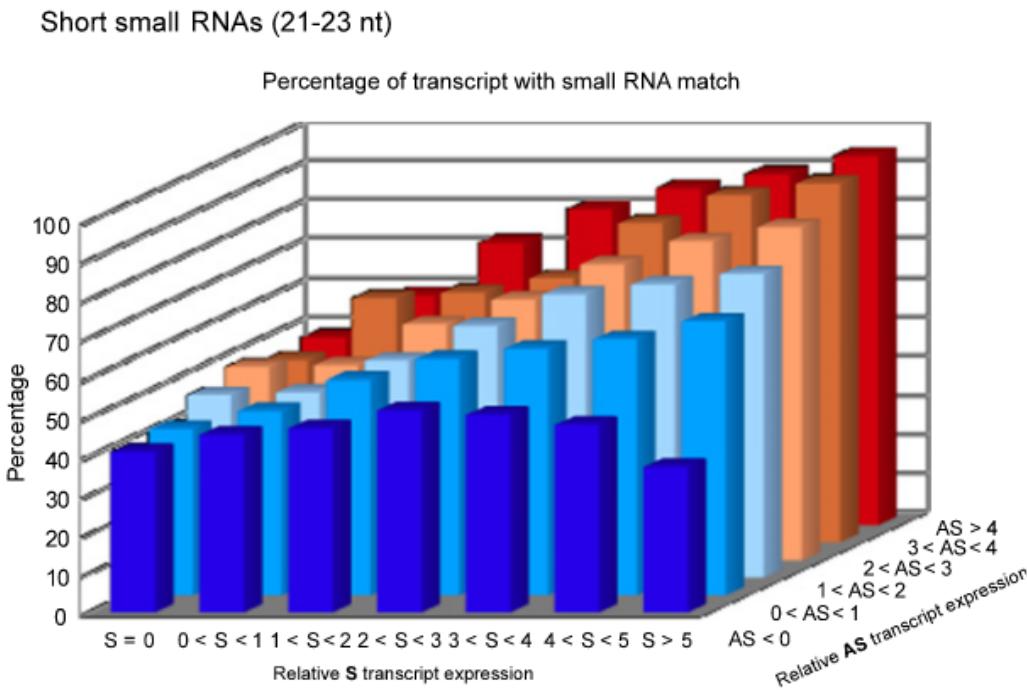


FIGURE 3.3 – Pourcentage de transcrits correspondants avec des petits-ARN. Les petits-ARN sont d'une taille comprise entre 21 et 23 nucléotides. Les pourcentages sont indiqués en fonction de l'expression (en log) des transcrits sens (S) et anti-sens (AS). Graphique issu de [Celton *et al.*, 2014].

Nous avons ainsi extrait des échantillons de l'atlas pour former un jeu de données associé à la maturation du fruit et composé de 22 échantillons de fruit à la récolte (Harvest – H) et 22 échantillons de fruit 60 jours après la récolte (60 Days After Harvest – 60DAH). À chaque échantillon de H correspond un échantillon de 60DAH. Les données étant issues de plusieurs projets différents, nous récupérons les données brutes afin de leur appliquer le même pré-traitement. La table 3.1 liste les échantillons utilisés avec le numéro d'accès dans la base de données Gene Expression Omnibus (<https://www.ncbi.nlm.nih.gov/geo/>) permettant d'accéder publiquement aux données. La particularité de ces données est que plusieurs génotypes de pomme différents sont utilisés, c'est-à-dire que nous utilisons différentes variétés de pommes. Notre étude permet donc de mettre en valeur des mécanismes qui sont communs à l'ensemble des variétés utilisées.

Les données brutes de transcriptomique doivent subir un certain nombre de pré-traitements et notamment une étape de normalisation des données. Classiquement, deux normalisations sont effectuées : une normalisation intra-puce et une normalisation inter-puce. La normalisation intra-puce est effectuée grâce à la méthode Lowess afin d'annuler différents effets techniques. La normalisation inter-puce s'effectue par la méthode de soustraction du bruit de fond. Lorsque les données d'intensités après ces deux normalisations sont représentées par échantillon, on observe une disparité dans la répartition de ces intensités. Afin d'étudier le comportement d'un ensemble de sondes sur l'ensemble des expériences, il semble nécessaire de faire une normalisation des données par échantillon.

La normalisation des données par la méthode de quantiles permet de faire cette harmonisation des

TABLE 3.1 – Liste des échantillons utilisés pour former les expériences H et 60DAH. Le numéro d'accession donne accès aux données publiées dans la base de données Gene Expression Omnibus (<https://www.ncbi.nlm.nih.gov/geo/>).

H	60DAH	Accession
M48_Harvest	M48_60DAH	GSE59947
M16_Harvest	M16_60DAH	
M20_Harvest	M20_60DAH	
M74_Harvest	M74_60DAH	
M49_Harvest	M49_60DAH	
M40_Harvest	M40_60DAH	
I095-P12-rec	I095-PH-2mois	GSE64079
H097-P12-rec	H097-PH-2mois	
H074-P12-rec	H074-PH-2mois	
I016-PH-rec	I016-P12-2mois	
V083-PH-rec	V083-P12-2mois	
I062-PH-rec	I062-P12-2mois	
V034-PH-rec	V034-PH-2mois	
W029-PH-rec	W029-PH-2mois	
Ari_re_2012	Ari_M2_2012	GSE101716
Gre_re_2012	Gre_M2_2012	
Fuj_re_2011	Fuj_M2_2011	
Jon_re_2011	Jon_M2_2011	
Fuj_re_2012	Fuj_M2_2012	
Jon_re_2012	Jon_M2_2012	
Gal_re_2012	Gal_M2_2012	
Cox_re_2012	Cox_M2_2012	

données par échantillon [Bullard *et al.*, 2010, Qiu *et al.*, 2013]. Pour éviter de superposer plusieurs normalisation de même type qui ajouterait de la variance dans les résultats, la normalisation par quantiles est faite suite à la normalisation intra-puce et en remplacement de la normalisation par soustraction du bruit de fond (figure 3.4).

Nous avons effectué la normalisation par les quantiles pour l'ensemble des 44 échantillons, puis nous avons formé les deux groupes d'échantillons pour chacune des expériences. À l'issue de la normalisation par les quantiles, nous obtenons des valeurs d'intensité comprises entre 7 et 16 log. Afin de considérer qu'une sonde n'est pas exprimée lorsqu'elle avoisine 0 log d'expression, nous effectuons une soustraction du bruit de fond de 7 log pour toutes les sondes. Grâce à cela, nous pouvons considérer qu'une sonde est exprimée dès lors que l'expression mesurée dépasse 1 log.

La molécule de rétrotranscriptase qui permet, à partir des transcrits, de produire l'ADN complémentaire qui se fixe sur les sondes de la puce, ne s'arrête pas toujours au bout du brin d'ARN qu'elle copie et quelquefois continue en repartant dans l'autre sens [Perocchi *et al.*, 2007]. Dans ce cas, cela crée un bruitage d'expression d'anti-sens en provenance d'un transcrit sens. Pour supprimer cet artefact, nous sommes plus restrictif pour les sondes anti-sens où nous considérons qu'une sonde anti-sens est exprimée si elle dépasse le seuil de 2 log. Pour simplifier le traitement, nous soustrayons 1 log supplémentaire pour les sondes anti-sens, et ainsi nous considérons qu'une sonde sens ou anti-sens est exprimée si son niveau d'expression dépasse 1 log. Dans les cas où un niveau d'intensité ainsi obtenu serait négatif, il est

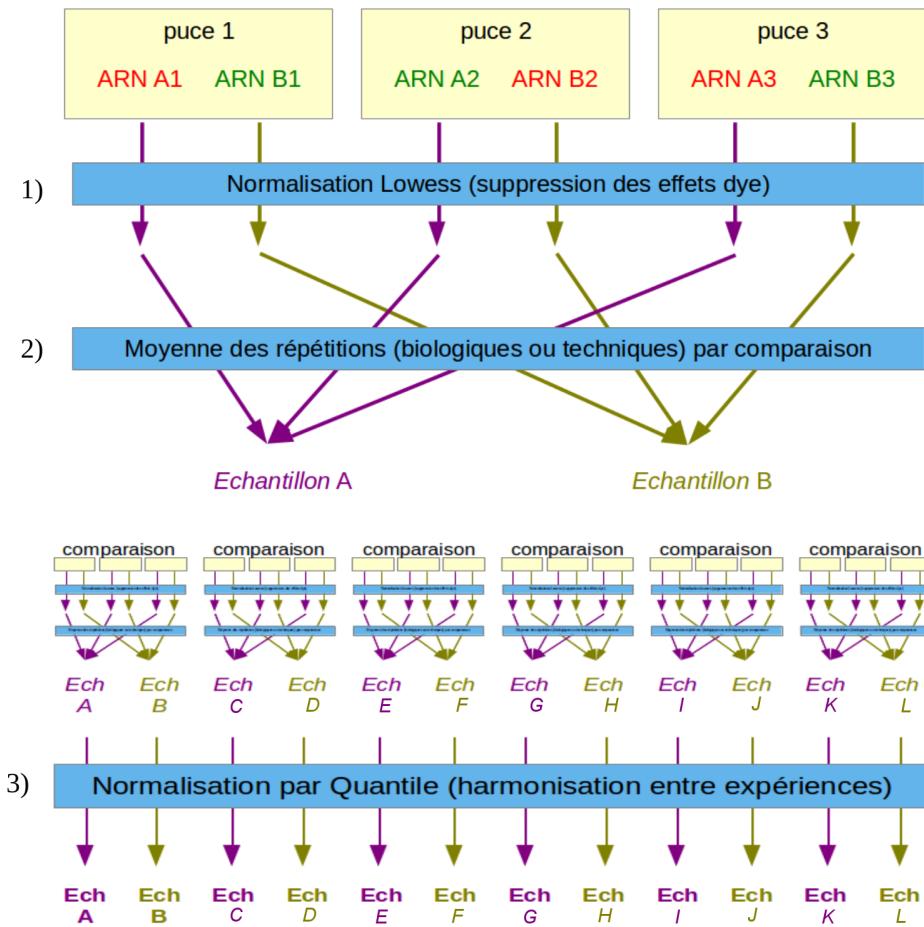


FIGURE 3.4 – Étapes de normalisation des échantillons en utilisant la normalisation par les quantiles. La normalisation s'effectue en trois étapes. La première étape est une normalisation intra-puce grâce à la méthode Lowess. La deuxième étape consiste à créer un seul échantillon à partir des répétitions en prenant la moyenne des répétitions. La troisième étape est une normalisation inter-échantillon grâce à la méthode des Quantiles.

ramené à 0. Après cette soustraction, nous obtenons des données d'intensités normalisées entre 0 et $9 \log$.

Nous n'allons pas travailler avec l'ensemble du génome, mais nous allons nous intéresser aux transcrits qui évoluent entre l'expérience H et 60DAH. L'*évolution* d'un transcrit (définition 3.1) est une discrétilisation de la différence du niveau d'expression moyen d'un transcrit entre deux conditions expérimentales.

Définition 3.1 (Évolution). Soient un transcrit p , les niveaux d'expression X_1 et X_2 de p respectivement dans les échantillons de la première et de la seconde expérience, et t un seuil d'expression.

L'évolution $e_t(p)$ d'un transcrit p est telle que :

$$e_t(p) = \begin{cases} -1 & \text{si } \overline{X_2} - \overline{X_1} \leq -t \\ 0 & \text{si } -t < \overline{X_2} - \overline{X_1} < t \\ 1 & \text{si } t \leq \overline{X_2} - \overline{X_1} \end{cases}$$

Nous définissons les transcrits différentiellement exprimés entre les deux expériences comme des *transcrits d'intérêt* (définition 3.2). Un transcript d'intérêt est un transcript qui possède une évolution non-nulle entre les deux expériences.

Définition 3.2 (Transcrit d'intérêt). *Soient un transcript p présent dans deux expériences, et t un seuil d'expression.*

Le transcript p est un transcript d'intérêt si, et seulement si :

$$e_t(p) \neq 0$$

3.4 Statistiques descriptives

Dans cette section, nous décrivons les données utilisées, nous commençons par décrire l'ensemble des données, à l'échelle du génome complet. Les données utilisées ici sont après normalisation par les quantiles et après la suppression du bruit de fond.

La figure 3.5 et la figure 3.6 montrent la répartition des valeurs d'intensités pour chacun des échantillons respectivement de l'expérience H et 60DAH. Sur chacune des figures, le premier graphe est réalisé à partir de l'ensemble des données, alors que les graphes situés en-dessous sont réalisés à partir des données sens uniquement (à gauche), et anti-sens uniquement (à droite). On remarque que la normalisation par les quantiles force chacun des échantillons à avoir la même médiane. La distribution des log-intensités sur la figure 3.7 pour l'expérience H et sur la figure 3.8 pour l'expérience 60DAH montre bien également que la majorité des sondes ne sont pas différentiellement exprimées.

On remarque que le niveau d'expression moyen des transcrits anti-sens est plus faible que celui des transcrits sens (figure 3.9). Cela peut s'expliquer par la suppression du bruit de fond qui supprime 1 log de plus pour les transcrits anti-sens, mais on observe sur la figure 3.9 que la différence des niveaux moyens d'expression entre les transcrits sens et anti-sens est supérieure à 1 log.

La figure 3.10 fait apparaître le nombre de transcrits sens exprimés, le nombre de transcrits anti-sens exprimés et le nombre de couples sens/anti-sens exprimés. On remarque que quelque soit le seuil utilisé, il y a une très faible différence entre le nombre de transcrits anti-sens exprimés et le nombre de couples exprimés. Cela s'explique par le fait que dans tous les cas il y a moins d'anti-sens exprimés que de sens (confirmé récemment par des données pommier RNA-Seq orienté produites par l'IRHS où le ratio du niveau d'expression anti-sens sur le niveau d'expression sens est de 1 pour 10), c'est pourquoi le nombre de couples exprimés est plus proche du nombre d'anti-sens que du nombre de sens. Mais ce faible écart montre aussi que les anti-sens exprimés sont pour la plupart des anti-sens complémentaires de sens qui sont également exprimés. On retrouve donc plus souvent un transcript sens sans son anti-sens complémentaire, par contre un transcript anti-sens est presque toujours en présence de son sens complémentaire.

Nous regardons maintenant les sondes qui évoluent entre les deux expériences H et 60DAH.

Communément, le seuil de 1 log est utilisé pour déterminer si un transcript est exprimé ou s'il est

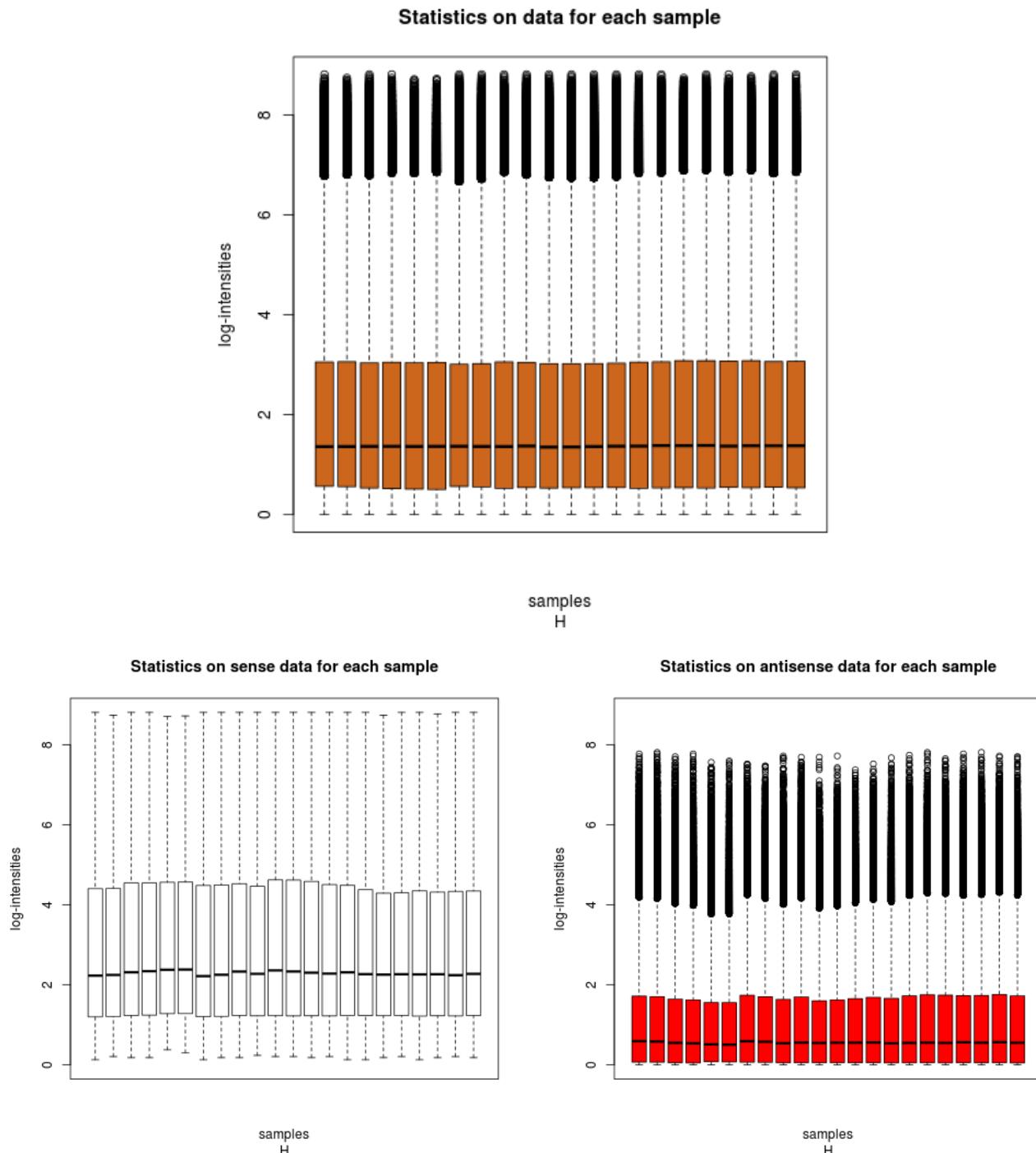


FIGURE 3.5 – Boxplots des log-intensités pour chacun des 22 échantillons de l’expérience H. Le graphique du haut a été fait à partir de l’ensemble des données sens et anti-sens. Le graphique en bas à gauche concerne les données sens uniquement, et le graphique en bas à droite les données anti-sens uniquement. Dans tous les cas, les données sont après normalisation par les quantiles et suppression du bruit de fond. Ces box plots indiquent les valeurs minimum et maximum à l’extrémité des moustaches, le premier et le troisième quartile à l’extrémité de la boîte, et la médiane est représentée par la barre dans la boîte. Les valeurs anormalement éloignées du reste des données (cercles noirs) ne sont pas considérés dans les calculs de la box plot.

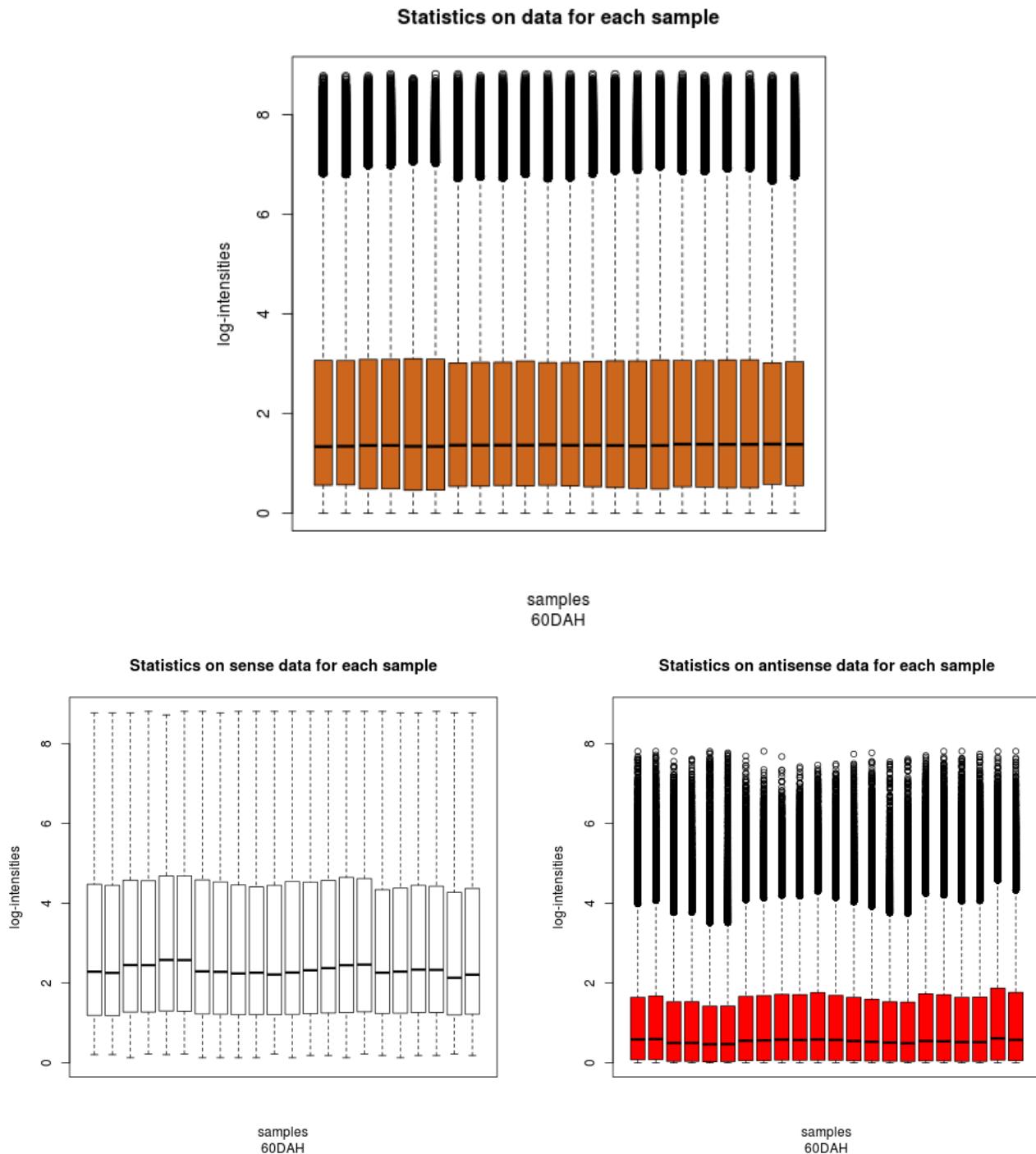


FIGURE 3.6 – Boxplots des log-intensités pour chacun des 22 échantillons de l’expérience 60DAH. Le graphique du haut a été fait à partir de l’ensemble des données sens et anti-sens. Le graphique en bas à gauche concerne les données sens uniquement, et le graphique en bas à droite les données anti-sens uniquement. Dans tous les cas, les données sont après normalisation par les quantiles et suppression du bruit de fond.

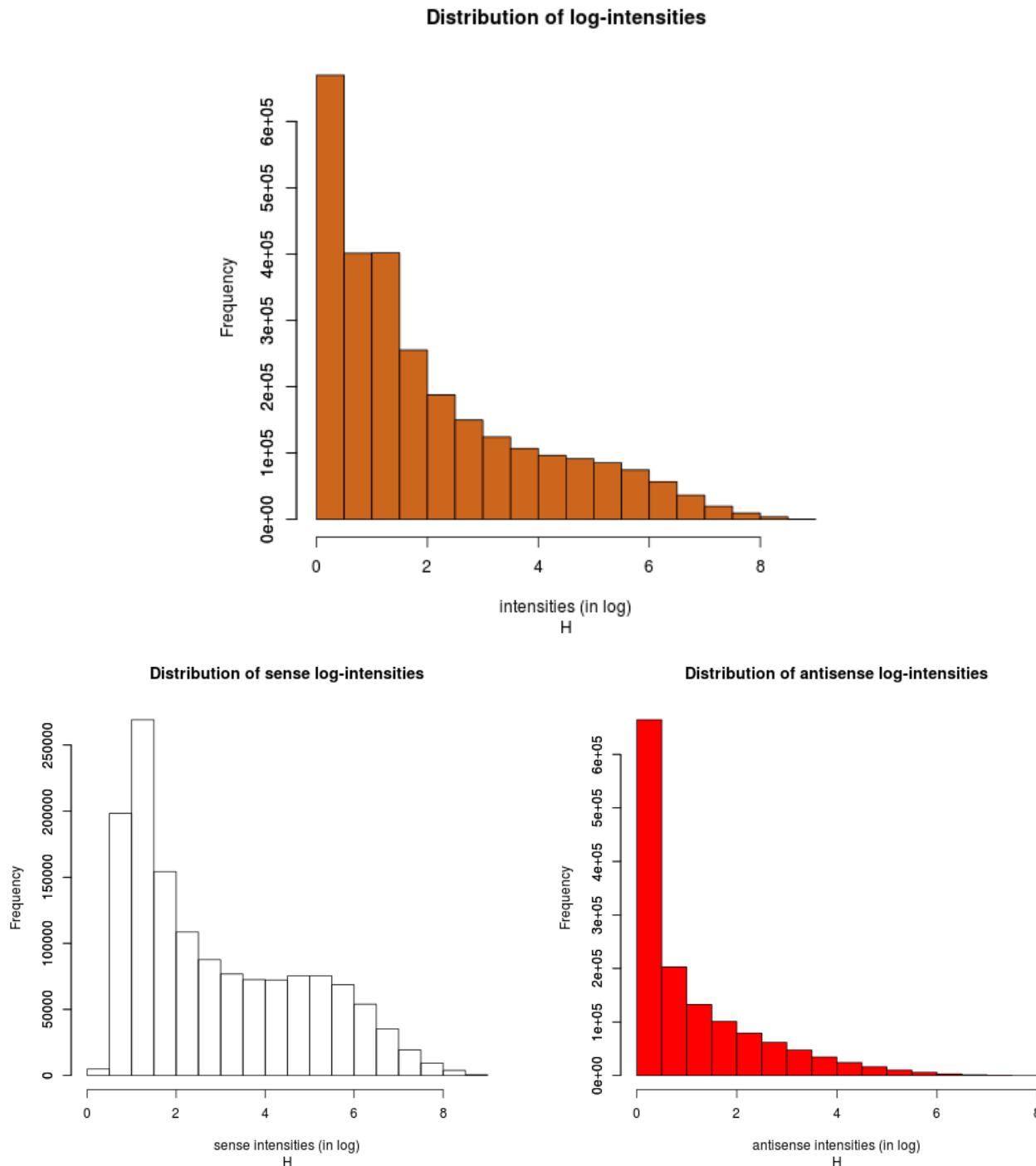


FIGURE 3.7 – Distribution des log-intensités pour chacun des 22 échantillons de l’expérience H. Le graphique du haut a été fait à partir de l’ensemble des données sens et anti-sens. Le graphique en bas à gauche concerne les données sens uniquement, et le graphique en bas à droite les données anti-sens uniquement. Dans tous les cas, les données sont après normalisation par les quantiles et suppression du bruit de fond.

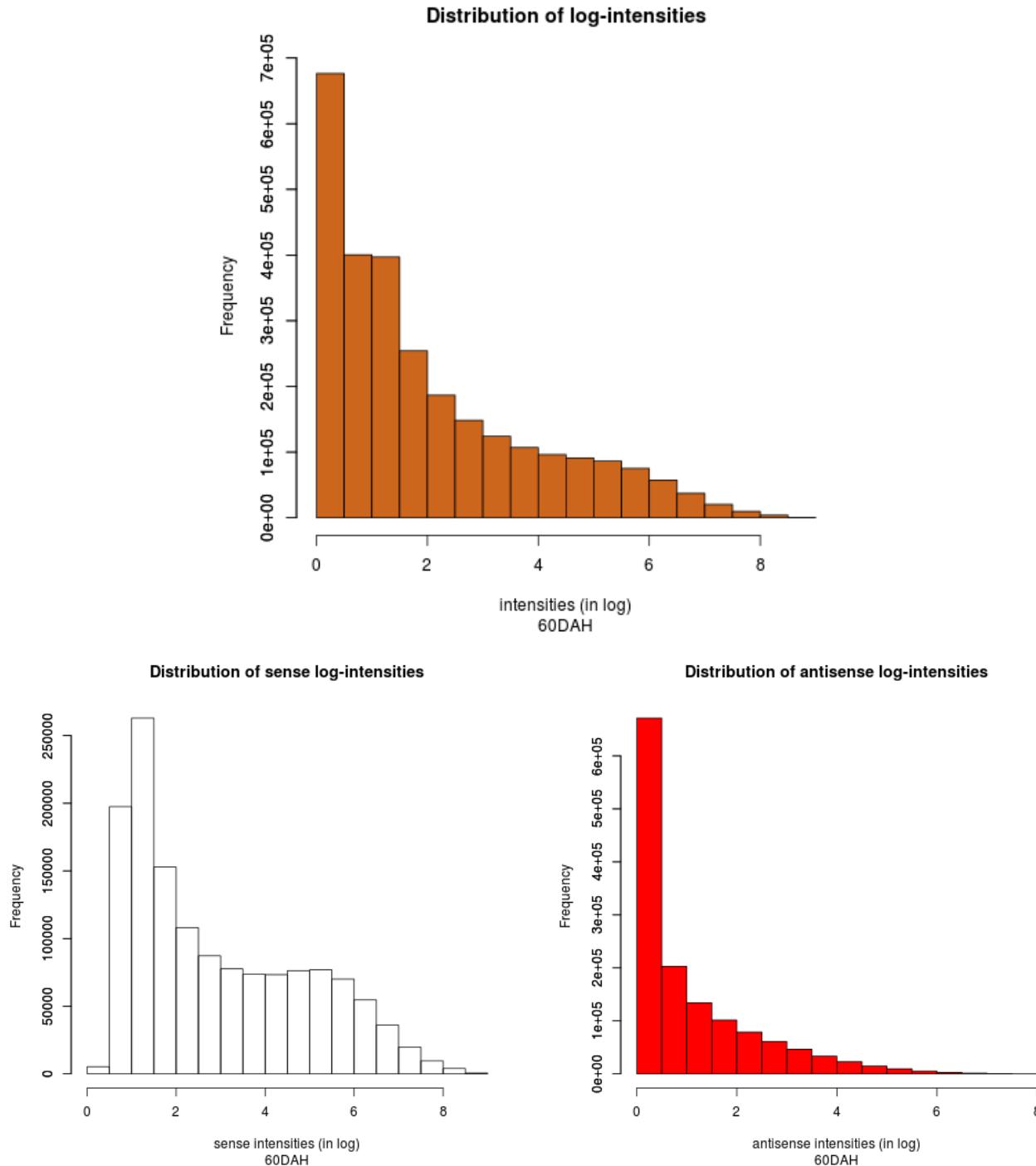


FIGURE 3.8 – Distribution des log-intensités pour chacun des 22 échantillons de l’expérience 60DAH. Le graphique du haut a été fait à partir de l’ensemble des données sens et anti-sens. Le graphique en bas à gauche concerne les données sens uniquement, et le graphique en bas à droite les données anti-sens uniquement. Dans tous les cas, les données sont après normalisation par les quantiles et suppression du bruit de fond.

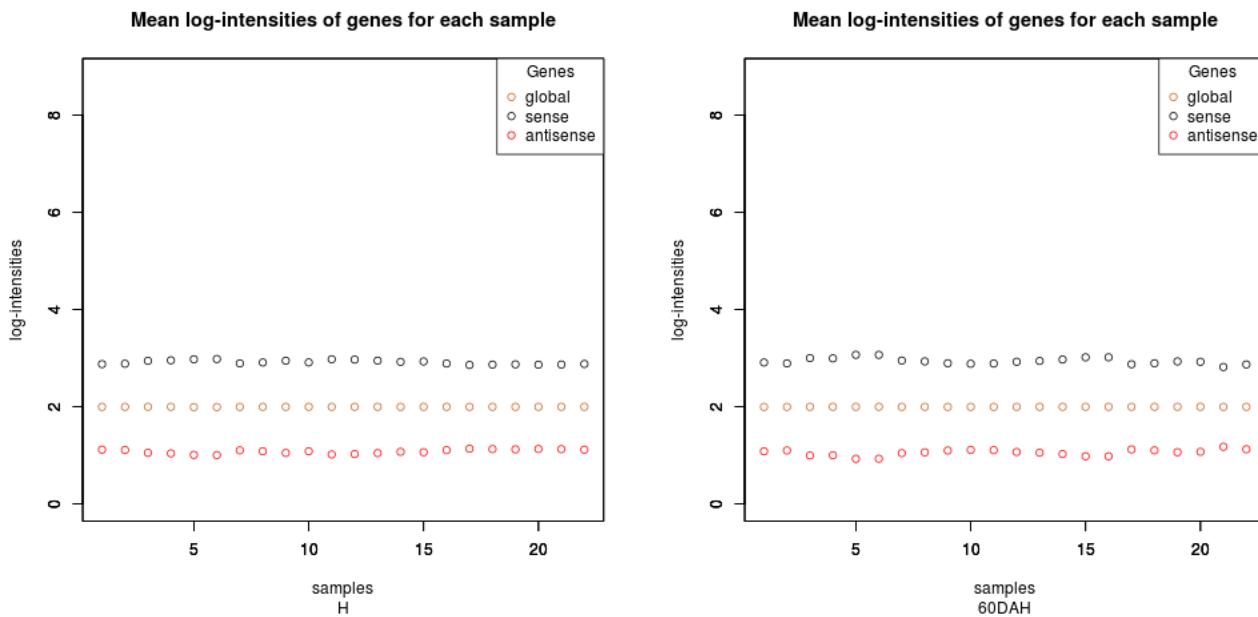


FIGURE 3.9 – Moyennes des log-intensités pour chacun des 22 échantillons des expériences H (gauche) et 60DAH (droite). Les données sont après normalisation par les quantiles et suppression de 7 log pour les données sens, et de 8 log pour les données anti-sens. Les points oranges sont pour les données sens et anti-sens, les points noirs pour les données sens uniquement, et les points rouges pour les données anti-sens uniquement.

différentiellement exprimé. La figure 3.11 montre le nombre de transcrits, sens d'un côté et anti-sens de l'autre, différentiellement exprimés entre les expériences H et 60DAH en fonction de trois seuils différents. Nous notons qu'un seuil de 1 log permet d'obtenir un nombre de transcrits avec lequel il est possible d'inférer des réseaux. Le seuil de 1 log pour définir les transcrits d'intérêt ainsi que l'évolution d'une sonde a donc été fixé.

Nous avons indiqué que chaque sonde est associée à une spécificité de 1 à 4 indiquant la confiance entre l'expression mesurée et le transcript ciblé. Les sondes de spécificité 4 peuvent s'hybrider avec un transcript qui est totalement différent de celui ciblé. Après avoir normalisé les données avec l'ensemble des données, nous souhaitons voir l'effet du retrait des sondes 4 sur l'évolution des sondes.

Nous utilisons donc le seuil de 1 log pour définir l'évolution et les gènes d'intérêt. La table 3.2 montre l'évolution des couples de sondes, dans la globalité et sans les sondes 4. On peut ainsi voir comment évoluent les sondes, mais également les couples de sondes sens/anti-sens. Globalement, entre H et 60DAH, on remarque que les sondes qui évoluent le font principalement de manière négative, c'est-à-dire que le niveau d'expression des gènes est plus bas deux mois après la récolte. On remarque également de manière surprenante qu'il est rare qu'un sens et son anti-sens complémentaire évoluent de manière opposée : seuls quatre couples sont dans ce cas. Il y a également une proportion plus importante de transcripts sens qui ont un niveau d'expression plus élevé dans 60DAH que dans H : environ 40% des transcripts sens ont une évolution positive contre seulement 26% pour les transcripts anti-sens.

Nous avons donc décidé de ne pas prendre en compte ces sondes dans notre étude afin d'être le

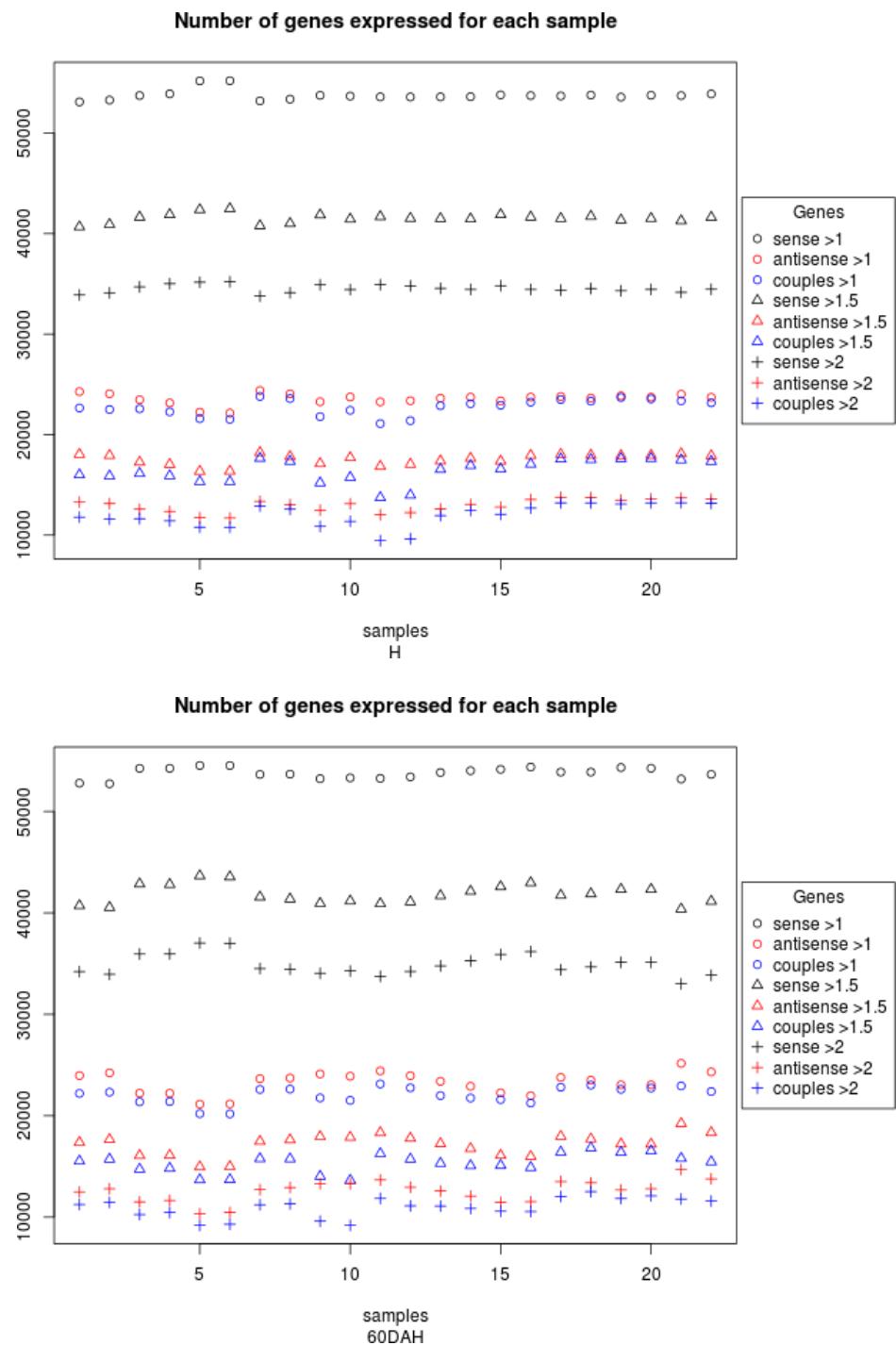


FIGURE 3.10 – Nombre de gènes exprimés selon un seuil pour chaque échantillon de H (haut) et 60DAH (bas). Les ronds sont pour un seuil de 1 log, les triangles pour un seuil de 1.5 log et les croix pour un seuil de 2 log. En noir les transcrits sens, en rouge les transcrits anti-sens et en bleu les couples, c'est-à-dire à la fois le transcript sens et le transcript anti-sens complémentaire ont un niveau d'expression qui dépasse le seuil.

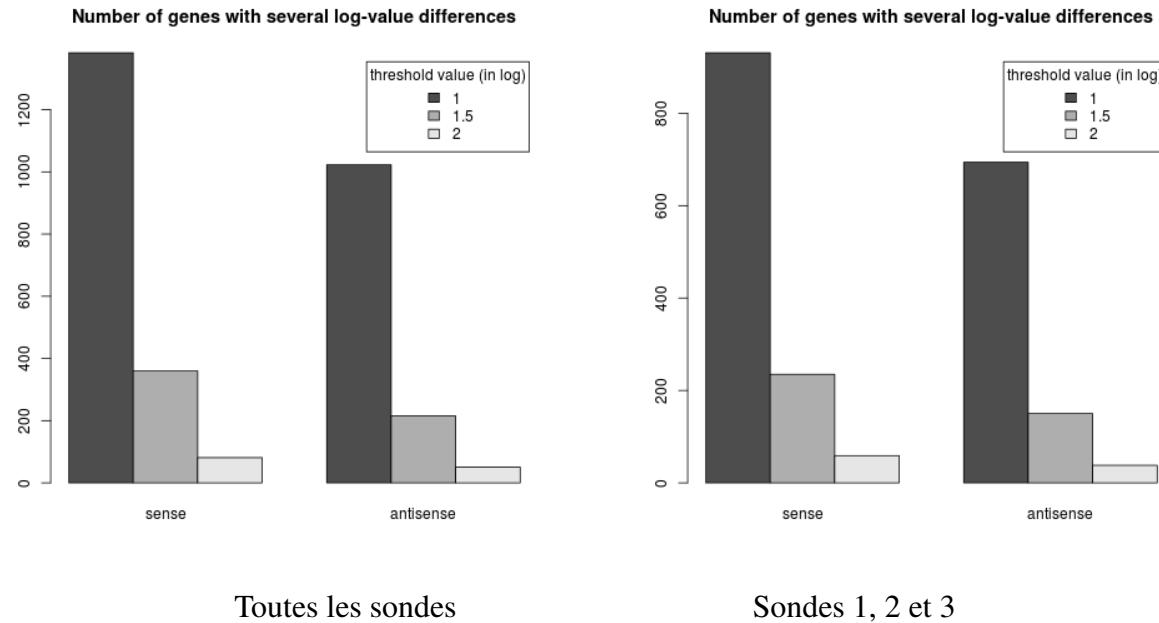


FIGURE 3.11 – Nombre de transcrits sens (groupe de gauche) et anti-sens (groupe de droite) différemment exprimés entre H et 60DAH selon un seuil. Le graphique de gauche concerne l'ensemble des sondes. Le graphique de droite concerne les sondes avec une spécificité de 1, 2 ou 3 ; les sondes 4 sont exclues.

TABLE 3.2 – Évolutions des couples de sondes. Cette table dénombre les sondes sens (S) et anti-sens (AS) en fonction de leur évolution (-1, 0, 1). La table de gauche concerne l'ensemble des sondes. La table de droite concerne seulement les sondes avec une spécificité de 1, 2 ou 3 ; les sondes 4 sont donc retirées.

		S		
		-1	0	1
AS		-1	220	544
		0	645	60 908
		1	4	176
				77

Toutes les sondes

		S		
		-1	0	1
AS		-1	136	371
		0	420	46 635
		1	3	123
				60

Sondes 1, 2 et 3

plus précis possible dans l’interprétation biologique des résultats. Dans la suite des travaux, les données utilisées sont les données normalisées, supprimées du bruit de fond, dont la spécificité des sondes est comprise entre 1 et 3, et restreintes aux transcrits d’intérêt à 1 log.

En utilisant les sondes dont la spécificité est comprise entre 1 et 3, nous dénombrons 96 120 sondes. La moitié de ces sondes sont des sondes sens, l’autre moitié de ces sondes sont les anti-sens complémentaires. Avec le seuil de 1 log utilisé pour définir un transcrit d’intérêt sur ces données, nous obtenons 1 625 transcrits d’intérêt : 931 transcrits sens et 694 transcrits anti-sens.

Ces 1 625 transcrits d’intérêt forment 1 425 couples de transcrits sens/anti-sens. Parmi ces 1 425 couples on en dénombre ainsi 200 pour lesquels le transcrit sens et le transcrit anti-sens sont des transcrits d’intérêt. Les 1 225 couples restants sont donc des couples pour lesquels soit le transcrit sens, soit le transcrit anti-sens, est un transcrit d’intérêt. Concernant l’évolution des 200 couples formés de deux transcrits d’intérêts complémentaires, 196 voient une évolution identique de leurs transcrits, et donc seulement quatre couples voient une évolution opposée entre le transcrit sens et le transcrit anti-sens. On peut voir l’évolution de ces quatre couples dans la table 3.2 : trois d’entre eux ont une évolution négative de leur transcrit sens et une évolution positive de leur transcrit anti-sens, et un seul couple a son transcrit sens qui évolue positivement tandis que son transcrit anti-sens évolue négativement.

3.5 Première étude de réseau sur les données

Pour compléter notre étude descriptive des données, nous allons utiliser le logiciel WGCNA que nous avons présenté dans le chapitre 2. WGCNA construit un réseau de gènes basé sur la corrélation linéaire et propose différents moyens pour analyser ce réseau. Nous l’avons donc utilisé pour analyser les réseaux que nous obtenons sur les données pommier.

L’objectif de cette étude est de voir si les transcrits anti-sens ont un comportement particulier dans un réseau de gènes. Cette étude exploratoire nous permet également de voir comment se regroupent les transcrits anti-sens dans un réseau de gènes lorsqu’on les intègre dans l’inférence de réseau. WGCNA a l’avantage de regrouper les gènes du réseau en modules. Nous pouvons donc observer, sur nos données, si les transcrits anti-sens se regroupent dans un seul module, ou si les modules sont composés à la fois de transcrits sens et anti-sens.

Nous avons donc utilisé WGCNA sur les données de l’expérience 60DAH, avec les données sens et anti-sens. La figure 3.12 montre le dendrogramme obtenu et la répartition des gènes dans les modules. Sur ce dendrogramme on peut observer 13 modules en plus du module gris. Ce module gris regroupe tous les gènes qui n’ont pu être affectés dans un module présentant une forte connectivité. Lors de la coupe du clustering hiérarchique de WGCNA, un module trop petit ne sera pas considéré, et les gènes de ce module seront donc placés dans le module gris. Ici nous avons utilisé une taille minimum de 20 gènes par modules.

Les 1 625 transcrits utilisés sont composés de 931 transcrits sens et 694 transcrits anti-sens. Parmi ces

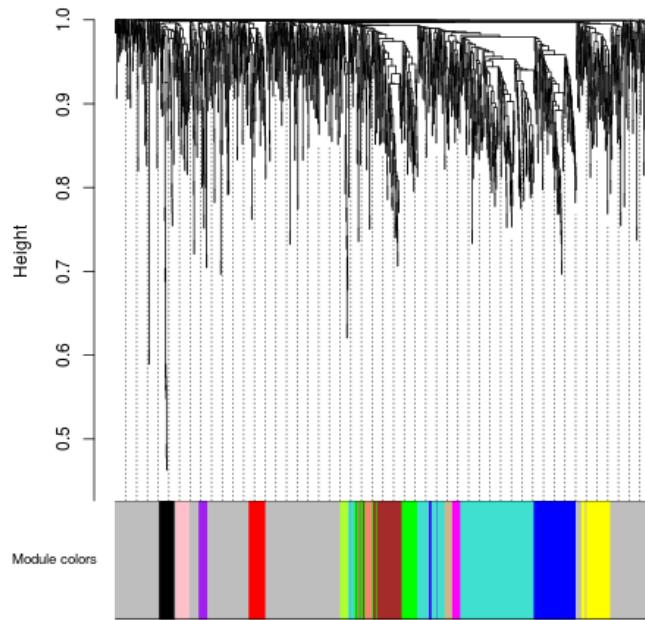


FIGURE 3.12 – Dendrogramme des gènes de l’expérience 60DAH. Le module dans lequel un gène a été assigné est indiqué sous le dendrogramme avec la couleur du module correspondant.

1 625 transcrits, 669 transcrits, dont 418 transcrits sens, n’ont pu être assignés à un module spécifique et sont donc regroupés dans le module gris (table 3.3). Nous observons donc que, même si beaucoup de transcrits ne sont pas affectés à un module, ce ne sont pas que les transcrits anti-sens qui ont été mis de côté par la méthode. Ainsi en intégrant, dans la méthode d’inférence, les transcrits anti-sens de la même manière que les transcrits sens, nous remarquons qu’une co-expression entre des transcrits sens et des transcrits anti-sens est détectée. Rappelons que seules les corrélations linéaires sont considérées par WGCNA.

Lorsque nous observons la répartition des transcrits anti-sens dans les modules (table 3.3), nous pouvons remarquer qu’il n’existe pas un module spécifique aux transcrits sens ou anti-sens. Dans chaque module, on observe la présence de transcrits sens et de transcrits anti-sens. On observe également que la répartition des transcrits dans les modules ne correspond pas à la distribution générale : le ratio sens/anti-sens varie selon les modules.

Cette étude exploratoire sur l’intégration des transcrits anti-sens dans une méthode d’inférence de réseaux de gènes nous confirme qu’il est possible de traiter les transcrits anti-sens comme les transcrits sens. Vu la répartition des transcrits dans les modules détectés par WGCNA, nous pouvons également dire que les transcrits anti-sens sont co-exprimés avec une majorité des transcrits d’intérêt sens, et non pas seulement un petit groupe d’entre eux.

Une analyse fonctionnelle sur chacun des modules a été réalisée. Nous avons reporté les résultats de cette étude en annexe B « Étude des données à partir de WGCNA » car les principes de l’analyse fonctionnelle vont être expliqués dans le chapitre 4.

TABLE 3.3 – Répartition des transcrits sens et anti-sens dans les modules. Il est indiqué la taille des modules, le nombre de transcrits sens (# S) et anti-sens (# AS) contenus dans chacun module.

60DAH				
Module	Taille module	# S	# AS	
grey	669	418	251	
turquoise	313	169	144	
brown	95	28	67	
red	50	31	19	
greenyellow	24	17	7	
green	72	27	45	
yellow	85	50	35	
blue	133	87	46	
pink	41	24	17	
magenta	27	17	10	
purple	26	11	15	
salmon	21	7	14	
black	46	26	20	
tan	23	19	4	



4

Analyse fonctionnelle différentielle

Dans une analyse transcriptomique, on obtient des données d'expression de gènes dans des conditions expérimentales déterminées. Ces données sont volumineuses et sont issues de tout ou partie d'un génome. Pour traiter ce volume de données, on sélectionne un ensemble de gènes par exemple grâce à une analyse de l'expression différentielle entre un contrôle et un traitement. L'une des premières étapes de l'interprétation biologique qui est réalisée est l'analyse fonctionnelle. L'analyse fonctionnelle permet d'identifier le rôle des gènes étudiés dans la cellule.

Nous rappelons d'abord le principe d'une analyse fonctionnelle ainsi que les outils d'analyse fonctionnelle utilisés pour la réaliser. Ensuite nous détaillons la méthode que nous proposons afin de réaliser une analyse fonctionnelle différentielle ; cette analyse a pour but de mettre en évidence des fonctions biologiques significativement représentées dans les données sens et anti-sens, alors qu'elles ne le sont pas dans les données sens seules. Nous donnons les résultats de l'analyse fonctionnelle pour les données pommier.

4.1 Gene Ontology	80
4.2 Analyse fonctionnelle d'un ensemble de gènes	82
4.2.1 Test d'enrichissement fonctionnel	82
4.2.2 Outils pour le test d'enrichissement fonctionnel	84
4.3 Tests d'enrichissement différentiels	86
4.3.1 Méthode	86
4.3.2 Résultats	89

4.1 Gene Ontology

Un des problèmes clés en biologie est de comprendre la fonction des gènes au sein des cellules. Les travaux et publications autour de ce sujet majeur sont nombreux, et les découvertes nouvelles doivent pouvoir être partagées par l'ensemble de la communauté. Pour cela, une première condition nécessaire est que les biologistes puissent tous parler de la même manière des fonctions étudiées. De plus, il a également été observé qu'une partie des protéines est partagée par tous les eucaryotes. Ainsi, en étudiant le rôle d'une protéine dans un organisme eucaryote, on peut transposer ce rôle à un autre organisme eucaryote partageant cette même protéine.

Le Gene Ontology Consortium est un consortium international qui a pour but de définir un vocabulaire unifié et commun à plusieurs organismes, afin de décrire les propriétés connues des gènes et des protéines qu'ils codent. Ce vocabulaire est organisé via un ensemble d'ontologies appelé Gene Ontology (GO) [Ashburner *et al.*, 2000]. De plus, ce consortium maintient également à jour les connaissances sur les gènes en proposant pour un ensemble d'organismes des annotations fonctionnelles des gènes, qui consistent à rattacher les gènes à un ou plusieurs termes du vocabulaire établi.

Une ontologie regroupe et hiérarchise un ensemble de termes rattaché à un domaine d'étude, elle correspond à un graphe orienté acyclique reliant les termes par des relations de subsomption (généralité). Les termes sont ainsi classés du plus spécifique au plus générique.

La figure 4.1 est un extrait d'une ontologie de la GO, où un noeud représente un terme de la GO et un lien représente la relation de subsomption où $t_1 \rightarrow t_2$ signifie que t_1 subsume t_2 ($t_1 \succ t_2$). Ainsi, on observe que le terme “cellular process” est plus générique que le terme “single-organisme cellular process”. Les termes les plus génériques sont appelés des termes de haut niveau, alors que les termes les plus spécifiques sont appelés des termes de bas niveau. Un terme peut être lié à plusieurs autres termes et peut ainsi avoir plusieurs termes plus spécifiques ou plus génériques. Dans la figure 4.1 le terme “single-organisme cellular process” a deux termes plus génériques : “cellular process” et “single-organisme process”, et ce dernier terme a quatre termes plus spécifiques.

La GO décrit les fonctions cellulaires suivant trois branches (ou ontologies) indépendantes :

- ‘biological process’ : Un terme de la GO ‘biological process’ fait référence à un processus biologique dans lequel le gène est impliqué. Un processus biologique s’effectue par l’association de plusieurs fonctions moléculaires via des transformations chimiques ou physiques. La GO ‘biological process’ est constituée de 3 163 termes. La figure 4.1 correspond à une partie de cette GO.
- ‘molecular function’ : L’activité biochimique du produit d’un gène est répertoriée dans la GO ‘molecular function’. Cela précise seulement l’activité en question sans dire où elle a lieu. Il n’est pas précisé non plus si le produit du gène fait cette activité seul ou dans un complexe protéique. La GO ‘molecular function’ est constituée de 2 284 termes.
- ‘cellular component’ : Enfin, la GO ‘cellular component’ indique où le produit d’un gène est actif. La GO ‘cellular component’ est constituée de 599 termes.

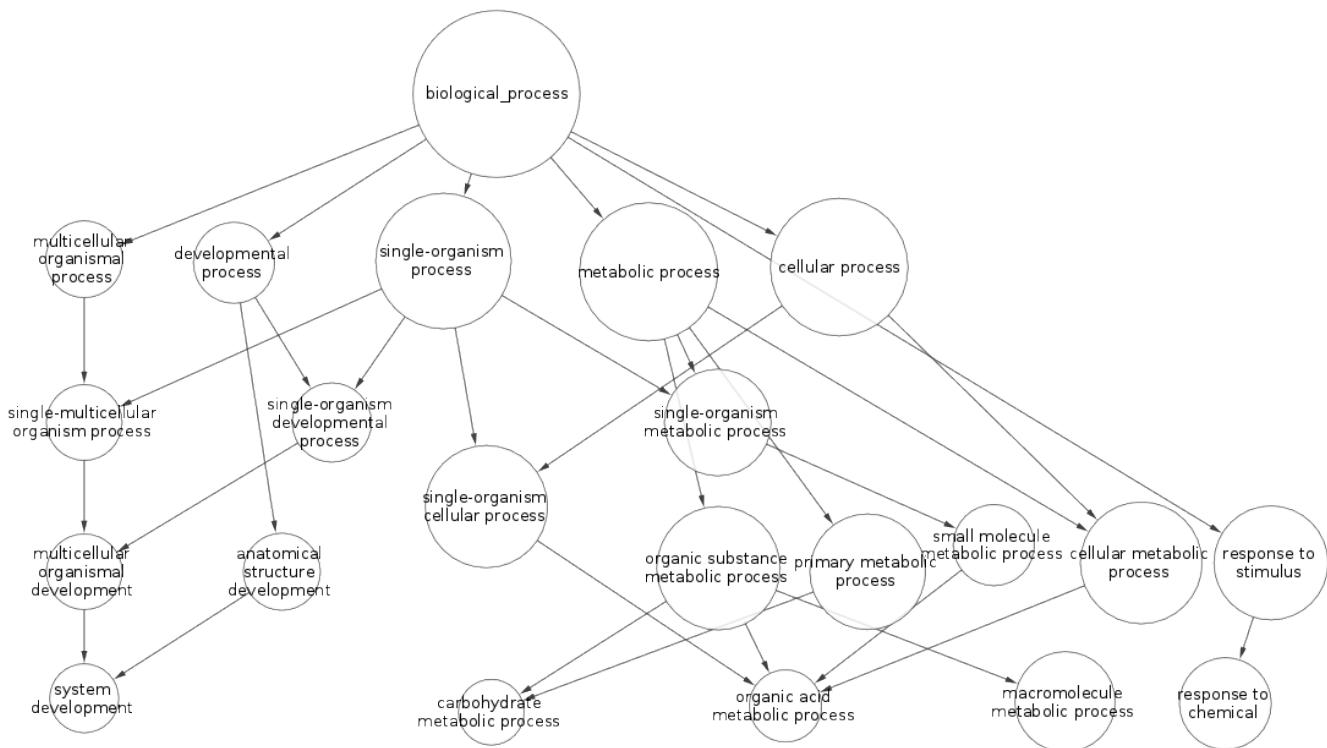


FIGURE 4.1 – Extrait de l'ontologie “Biological process”. Un nœud représente un terme de l'ontologie. Un lien représente la relation de subsomption entre deux termes : le terme source subsume le terme cible.

Dans la Gene Ontology, le terme le plus générique est celui qui porte le nom de l'ontologie.

La Gene Ontology est construite de telle manière qu'elle peut être appliquée à l'ensemble des organismes répertoriés. Cela signifie que dans les trois ontologies, tous les termes ne sont pas applicables pour tous les organismes. L'ensemble des termes définis est fait pour être le plus large possible et s'adapter au plus grand nombre d'organismes.

Le Gene Ontology Consortium regroupe l'ensemble des fichiers d'association de gènes avec les termes pour chacun des organismes étudiés et pour chacune des ontologies. Un fichier d'association, appelé fichier d'annotations, permet donc, entre autres, de lister les fonctions biologiques associées à un ensemble de gènes. Un gène peut ne pas posséder d'annotation pour une ontologie donnée, ou à l'inverse avoir plusieurs annotations. L'association entre un gène g et un terme t s'appuie en grande majorité sur des similarités de séquences de g avec un gène g' qui est déjà associé au terme t . La qualité d'association peut donc être très hétérogène.

Pour des études fonctionnelles avec des ensembles de gènes relativement petits, ou pour réaliser des études qui puissent donner une vue moins détaillée, le consortium met également à disposition une GO slim. Une GO slim est une version simplifiée de la GO complète, elle permet d'avoir une vue plus générale sur les fonctions représentées dans un ensemble de gènes. Seuls les termes les plus génériques composent la GO slim : il n'existe ainsi pas de relation de subsomption entre les différents termes de la GO slim.

TABLE 4.1 – Liste des termes de la GO slim ‘biological process’.

biological_process unknown	developmental processes	transport
signal transduction	cell organization and biogenesis	other cellular processes
DNA or RNA metabolism	protein metabolism	electron transport and energy pathways
transcription	other metabolic processes	response to abiotic and biotic stimulus
response to other stresses	other physiological processes	other biological processes

La GO slim ‘biological process’ est ainsi composée d’uniquement 15 catégories (table 4.1). Ces catégories englobent la totalité de la GO. Cela facilite ainsi les études en ne donnant que des grandes catégories de fonctions biologiques. Les catégories ont été choisies pour fournir une représentation la plus large possible de la distribution des fonctions biologiques des gènes. Ces catégories ont été réalisées principalement pour qu’un terme de la GO ne soit associé qu’à une seule catégorie de la GO slim. Ainsi les termes de la GO, et donc les gènes associés à ces termes, ne sont rattachés qu’à une seule catégorie de la GO slim. Seuls quelques termes de la GO se retrouvent associés à deux catégories de la GO slim. Par exemple, le terme “response to wounding” a deux parents dans la GO complète : “response to stress” et “response to external stimulus” ; dans la GO slim, les gènes associés à “response to wounding” seront ainsi associés à “response to stress” et à “other physiological processes”.

4.2 Analyse fonctionnelle d’un ensemble de gènes

Lors d’une analyse biologique, on peut identifier des ensembles de gènes qui ont un comportement similaire. On peut ainsi par exemple identifier les gènes qui sont différentiellement exprimés dans les conditions expérimentales étudiées ou on peut identifier des gènes qui, dans un réseau de gènes, forment un module. Lorsqu’un sous-ensemble de gènes est ainsi sélectionné, on peut imaginer de mener des recherches bibliographiques pour voir si des relations entre ces gènes ont déjà été identifiées. Grâce aux connaissances disponibles dans des bases de connaissances telles que la GO, on peut mener des études plus automatiques. Grâce aux annotations des gènes, on peut chercher quelles sont les fonctions biologiques associées aux gènes d’un tel ensemble. Plutôt que de simplement lister les fonctions, il est également possible de rechercher les fonctions biologiques qui sont sur-représentées dans un ensemble de gènes sélectionnés.

4.2.1 Test d’enrichissement fonctionnel

Le test d’enrichissement fonctionnel consiste à étudier si, dans un ensemble de gènes donné, une fonction biologique est plus représentée que dans un ensemble de gènes de même effectif tiré aléatoirement dans l’ensemble du génome. L’hypothèse nulle dans ce test est donc qu’il n’existe aucun lien entre les fonctions

identifiées et le mode de sélection de l'ensemble de gènes. Le test statistique réalisé est un test exact de Fisher qui se base sur la distribution de la loi hypergéométrique.

La loi hypergéométrique est une loi de probabilité discrète qui modélise le nombre de cas favorables dans un échantillon d'effectif donné, l'échantillon étant obtenu lors d'un tirage sans remise à partir d'une population composée de cas favorables et de cas non favorables, et le nombre de cas favorables dans la population étant connu.

La loi hypergéométrique \mathcal{H} est décrite par trois paramètres : la taille de la population N , la taille de l'échantillon n , et la probabilité p d'événement favorable dans la population. Lorsqu'une variable aléatoire discrète X suit la loi hypergéométrique $\mathcal{H}(N; n; p)$, le calcul de la probabilité $P(X = k)$ pour que l'événement favorable se produise k fois est :

$$P(X = k) = \frac{\binom{Np}{k} \binom{N(1-p)}{n-k}}{\binom{N}{n}} \quad (4.1)$$

Dans le cas d'un enrichissement fonctionnel, nous souhaitons tester la représentation d'une fonction f dans un échantillon de g gènes. La taille G du génome est connue, et la Gene Ontology nous permet de connaître le nombre E de gènes du génome qui sont étiquetés par la fonction f , ainsi que le nombre e de gènes de l'échantillon qui sont également étiquetés par f . L'équation 4.1 devient alors :

$$P(X = e) = \frac{\binom{E}{e} \binom{G-E}{g-e}}{\binom{G}{g}} \quad (4.2)$$

Le test d'enrichissement fonctionnel est un test de sur-représentation, on veut savoir quelle est la probabilité d'observer, dans un échantillon aléatoire de même taille que l'échantillon considéré, au moins autant de gènes étiquetés par la fonction f que ce qui est observé dans l'échantillon considéré. On calcule donc la probabilité d'avoir un nombre de gènes étiquetés par f supérieur ou égal à e , soit $P(X \geq e)$ (équation 4.3). Si cette probabilité est faible, on pourra conclure que la fonction f est sur-représentée dans l'ensemble de gènes considéré.

$$P(X \geq e) = 1 - \sum_{k=1}^{e-1} \frac{\binom{E}{k} \binom{G-E}{g-k}}{\binom{G}{g}} \quad (4.3)$$

Le test d'enrichissement fonctionnel est ainsi effectué pour chacune des fonctions représentées dans l'échantillon. Cela représente plusieurs centaines de tests pour la même analyse fonctionnelle. Lorsque plusieurs hypothèses sont testées, les p-valeurs doivent être corrigées pour pouvoir être comparées [Ge *et al.*, 2003]. Plusieurs méthodes peuvent être utilisées pour corriger les p-valeurs, telles que la correction de Bonferroni ou la correction de Benjamini et Hochberg. Cette correction permet de contrôler le taux de faux-positifs [Maere *et al.*, 2005].

La sur-représentation d'une fonction est déclarée statistiquement significative lorsque la p-valeur est inférieure à un seuil choisi, généralement égal à 0.05 ou 0.01. L'ensemble des gènes étudiés est alors

considéré comme « enrichi » par la fonction biologique trouvée.

4.2.2 Outils pour le test d'enrichissement fonctionnel

Pour réaliser un test d'enrichissement fonctionnel, les outils développés se basent sur la Gene Ontology (GO) pour les annotations des gènes et pour les ontologies. Puisque l'annotation associe à un gène l'ensemble des termes les plus spécifiques de l'ontologie, pour effectuer l'analyse fonctionnelle, les outils « déroulent » la hiérarchie de l'ontologie afin d'associer le gène avec l'ensemble de tous les termes qui le définissent. Ainsi le test d'enrichissement fonctionnel ne sera pas réalisé seulement avec les termes les plus spécifiques associés aux gènes étudiés, mais avec l'ensemble de tous les termes associés aux gènes étudiés. En reprenant l'ontologie de la figure 4.1, l'enrichissement fonctionnel de gènes annotés “single-organism cellular process” et “anatomical structure development” s’effectuera avec l’ensemble des termes suivants : “single-organism cellular process”, “cellular process”, “single-organism process”, “anatomical structure development”, “developmental process” et “biological_process”.

Displaying only results with P<0.05; click here to display all results						
	Arabidopsis thaliana (REF)	upload 1 (▼ Hierarchy NEW! ?)				
GO biological process complete	#	#	expected	Fold Enrichment	+/-	P value
flavonoid biosynthetic process	59	11	1.51	7.29	+	1.19E-03
↳ metabolic process	8027	274	205.43	1.33	+	3.99E-05
↳ flavonoid metabolic process	67	12	1.71	7.00	+	5.60E-04
response to light stimulus	566	35	14.49	2.42	+	5.58E-03
↳ response to radiation	585	35	14.97	2.34	+	1.13E-02
↳ response to abiotic stimulus	1491	79	38.16	2.07	+	2.99E-06
↳ response to stimulus	4612	193	118.03	1.64	+	2.15E-09
oxidation-reduction process	1223	67	31.30	2.14	+	1.67E-05
↳ single-organism metabolic process	2986	133	76.42	1.74	+	4.43E-07
↳ single-organism process	6744	261	172.59	1.51	+	1.85E-10
response to acid chemical	886	47	22.67	2.07	+	7.31E-03
↳ response to chemical	2084	106	53.33	1.99	+	3.26E-08
response to oxygen-containing compound	1144	58	29.28	1.98	+	2.12E-03
response to hormone	1200	60	30.71	1.95	+	2.13E-03
↳ response to organic substance	1466	72	37.52	1.92	+	3.22E-04
↳ response to endogenous stimulus	1208	60	30.92	1.94	+	2.62E-03
single-organism cellular process	3518	136	90.03	1.51	+	1.38E-03
↳ cellular process	8484	284	217.12	1.31	+	1.23E-04
Unclassified	8213	193	210.19	.92	-	0.00E00

FIGURE 4.2 – Capture d'écran de l'outil AmiGO de l'analyse fonctionnelle pour l'ensemble des transcrits sens.

Le Gene Ontology Consortium propose l'outil en ligne AmiGO [Carbon *et al.*, 2009] pour réaliser une analyse fonctionnelle. L'enrichissement fonctionnel est réalisé en fonction de l'ontologie sélectionnée, et le résultat de l'analyse est un tableau classé en fonction des p-valeurs et organisé selon la hiérarchie de l'ontologie. La figure 4.2 est une capture d'écran d'un résultat d'analyse fonctionnelle utilisant

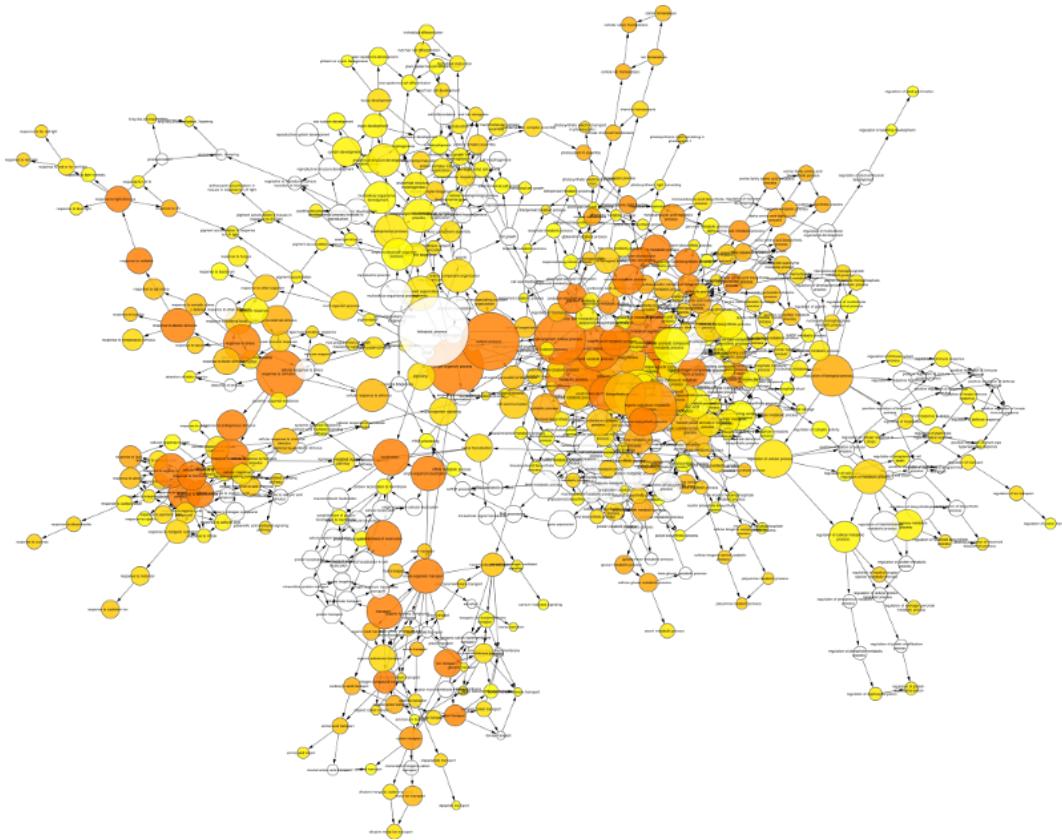


FIGURE 4.3 – Analyse fonctionnelle pour l’ensemble des transcrits sens obtenue grâce à BiNGO.

AmiGO. Seuls les résultats ayant une p-valeur inférieure à 0.05 sont affichés. Avec cette visualisation il est cependant difficile de bien observer les relations hiérarchiques entre certains termes. On ne voit pas, par exemple, à quels autres termes de la liste est relié le terme “response to oxygen-containing compound”. Il est également difficile de voir la hiérarchie entre les termes “flavonoid biosynthetic process”, “metabolic process” et “flavonoid metabolic process”.

L’outil BiNGO [Maere *et al.*, 2005] a été développé comme une App de Cytoscape. L’avantage de BiNGO par rapport à AmiGO est de permettre une restitution des résultats sous forme graphique, et avec les possibilités de navigation dans ce graphe offertes par Cytoscape. Le fonctionnement de BiNGO est comparable à celui d’AmiGO : on donne un ensemble de gènes et une ontologie, et on obtient une analyse fonctionnelle à la fois sous la forme d’une table, et d’un graphique (figure 4.3).

L’intérêt de BiNGO est donc dans la visualisation du résultat. La figure 4.3 est un résultat graphique d’analyse fonctionnelle de BiNGO. Le résultat est la visualisation de l’ontologie, où un nœud est étiqueté par un terme de la GO et un lien $t_1 \rightarrow t_2$ implique que le terme t_2 est une spécialisation du terme t_1 . Dans cette visualisation, la taille du nœud t est fonction du nombre de gènes associés au terme t dans l’ensemble de gènes de départ : plus le nombre de gènes associés à un terme est grand, plus le nœud sera gros. Les nœuds sont colorés par une échelle de couleurs allant du jaune au orange. La couleur indique la p-valeur associée au terme lors du test hypogéométrique : le jaune pour une p-valeur élevée et le orange pour une p-valeur faible.

Afin de dessiner un graphe connexe, BiNGO représente également les termes de l'ontologie qui ont des p-valeurs trop élevées mais qui permettent de relier les termes sur-représentés de la GO selon la hiérarchie. Ces nœuds sont visualisés avec une couleur blanche.

Dans la figure 4.3 on peut donc remarquer le plus gros nœud du graphe est “biological_process”, la racine de l'ontologie ‘Biological process’ : le terme est naturellement associé à l'ensemble de tous les gènes, et n'est donc pas significativement représentatif, mais il permet de dessiner l'ontologie et ainsi de hiérarchiser l'ensemble des termes.

En ce qui concerne l'utilisation de ces outils, AmiGO est plus simple à interroger que BiNGO. En effet, AmiGO étant en ligne et hébergé par le Gene Ontology Consortium, il suffit d'indiquer simplement quel organisme est étudié et quelle ontologie est désirée, puis l'outil va lui-même interroger la GO. De son côté, BiNGO ne peut pas interroger la GO en ligne, il faut donc lui fournir la GO, ainsi que le fichier d'association pour tous les gènes de l'organisme étudié et les termes de la GO.

4.3 Tests d'enrichissement différentiels

Comme on l'a dit plus haut, l'analyse fonctionnelle est souvent une première phase de questionnement lorsqu'une analyse bio-informatique a mis en évidence un ensemble de gènes parmi des données de transcriptomique. Dans l'étude que nous voulons mener, nos données transcriptomiques contiennent à la fois des transcrits sens associés à des gènes codants et des transcrits anti-sens dont on soupçonne qu'ils jouent un rôle de régulateurs sur certains transcrits sens dans certaines conditions. Nous proposons ici de mener une analyse fonctionnelle différentielle afin de mettre en évidence les fonctions biologiques concernées par l'action des transcrits anti-sens.

4.3.1 Méthode

Nous considérons ici que nous disposons à l'issue d'une analyse transcriptomique d'un ensemble de transcrits sens S et d'un ensemble de transcrits anti-sens AS , jugés potentiellement intéressants. Ces ensembles peuvent résulter, par exemple, d'une analyse différentielle de l'expression. L'analyse fonctionnelle différentielle se réalise en deux étapes. La première étape réalise deux test d'enrichissement fonctionnel : un test est effectué pour les données S contenant uniquement les transcrits sens, et un autre test est effectué pour les données $S \cup AS$ sens et anti-sens. La seconde étape est la comparaison des résultats de ces deux tests d'enrichissement. La figure 4.4 illustre le traitement des données lors de l'analyse fonctionnelle différentielle.

Deux analyses fonctionnelles sont donc réalisées sur les données. La première analyse est l'analyse que tout biologiste aurait fait : un enrichissement fonctionnel à partir de l'ensemble des gènes identifiés comme potentiellement intéressants. Afin de voir quelles informations peuvent apporter les données anti-sens, nous procédons à une seconde analyse fonctionnelle : l'enrichissement à partir de tous les transcrits considérés, sens et anti-sens. Nous souhaitons en effet étudier l'impact des nouveaux acteurs que sont

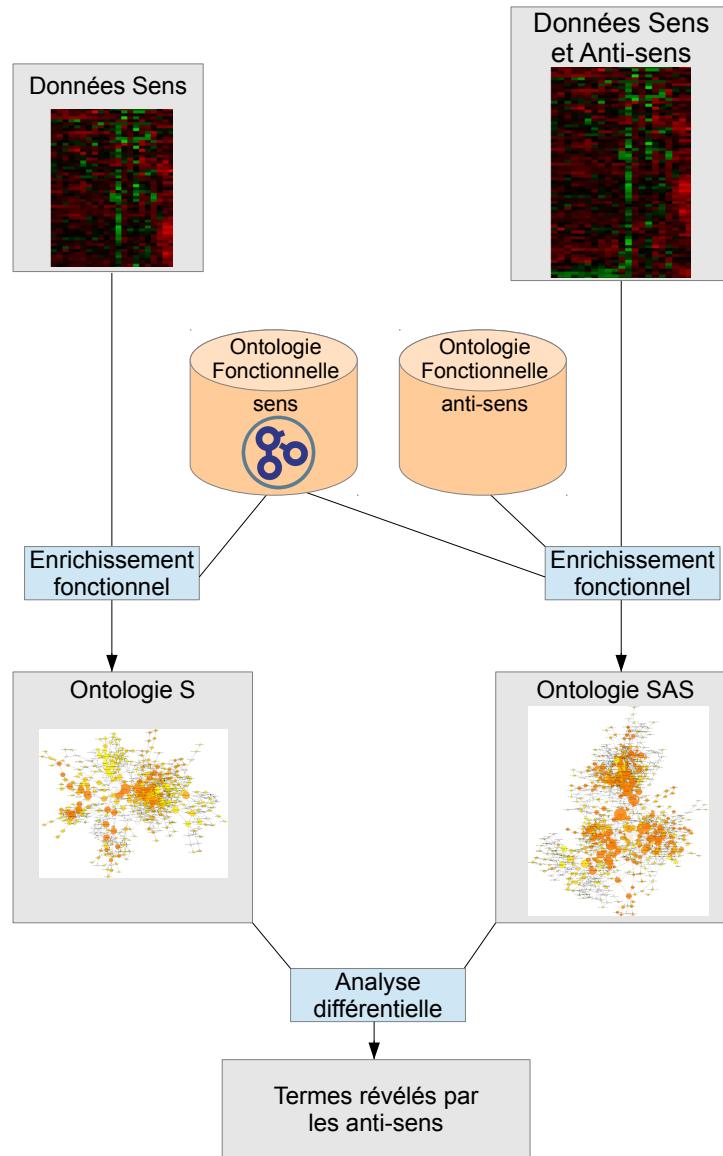


FIGURE 4.4 – Méthodologie de l'analyse fonctionnelle différentielle.

les anti-sens dans les analyses bio-informatiques classiquement réalisées. Nous proposons donc ici de considérer les transcrits anti-sens au même titre que les transcrits sens. Leur prise en compte modifie bien sûr le nombre de transcrits considérés et peut changer les résultats des tests d'enrichissement.

Puisque nous souhaitons connaître les processus biologiques impactés par l'intégration des données anti-sens, les tests d'enrichissement sont réalisés en interrogeant la GO ‘biological process’. Cependant il n'existe pas à l'heure actuelle d'annotations pour les transcrits anti-sens. Les familles multigéniques sont des ensembles de gènes qui ont des fonctions biochimiques similaires dans la cellule. Les membres d'une famille multigénique sont identifiés par leur similarité de séquence. Puisqu'un transcrit anti-sens a une séquence complémentaire du transcrit sens, il est raisonnable de considérer qu'il a un impact sur la même fonction biologique que son complémentaire et nous proposons donc d'annoter un transcrit anti-sens par le même terme que le transcrit sens complémentaire. De plus, parmi les rôles éventuels des transcrits anti-sens se trouve le mécanisme de PTGS (Post Transcriptional Gene Silencing) dans lequel le transcrit anti-sens s'hybride avec son sens complémentaire, qui va donc avoir un effet sur la fonction biologique assurée par ce gène.

Nous avons donc créé des annotations pour les transcrits anti-sens. Grâce à ces nouvelles annotations, nous pouvons réaliser le test d'enrichissement fonctionnel pour les données sens et anti-sens.

Il faut noter que chacune de ces analyses utilise donc un génome de référence (l'ensemble G de l'équation 4.3) différent. Si la GO contient G_S gènes annotés, l'analyse fonctionnelle pour les données sens est réalisée avec cet ensemble comme ensemble de référence. En revanche, pour l'analyse des données sens et anti-sens, les tests d'enrichissement s'effectuent avec un ensemble de référence G_{SAS} qui contient $2 \times G_S$ annotations. De la même manière, si pour une fonction f , la GO contient E gènes annotés par f , dans l'annotation complétée utilisée pour analyser les données sens et anti-sens, on aura $2 \times E$ transcrits annotés par f .

Il faut remarquer que dans les ensembles de transcrits intéressants que l'on veut analyser on ne trouve pas nécessairement un sens et son anti-sens associé. Les deux tests d'enrichissement peuvent donc conduire à des ensembles de fonctions biologiques, sur-représentées dans S d'une part, et dans $S \cup AS$ d'autre part, différentes. La seconde étape de l'analyse fonctionnelle différentielle consiste donc à comparer les listes de fonctions biologiques obtenues. Cette dernière étape nous permet d'identifier les *termes révélés par les anti-sens* (définition 4.1) c'est-à-dire les termes d'une ontologie qui n'auraient pas été identifiés comme sur-représentés en utilisant uniquement les données sens.

Définition 4.1 (Termes révélés par les anti-sens). *Soit une p-valeur seuil p_0 . Soient F_S l'ensemble des fonctions sur-représentées par l'ensemble des transcrits sens S dont la p-valeur est inférieure à p_0 , et F_{SAS} l'ensemble des fonctions sur-représentées par l'ensemble des transcrits sens S et anti-sens AS dont la p-valeur est inférieure à p_0 .*

L'ensemble des termes révélés par les anti-sens est l'ensemble des fonctions défini comme suit :

$$F_{SAS} \setminus F_S$$

L'ensemble des termes révélés par les anti-sens est calculé en faisant la différence entre les deux ensembles de termes significativement sur-représentés calculés lors des deux analyses fonctionnelles de la première étape. On part de l'ensemble des termes issus de l'analyse fonctionnelle contenant l'ensemble des données, auquel on retire tous les termes obtenus par l'analyse fonctionnelle des données sens uniquement.

4.3.2 Résultats

Nous avons réalisé l'analyse fonctionnelle différentielle avec les données pommier. Le pommier n'est pas un organisme recensé par le Gene Ontology Consortium, il n'existe donc pas d'annotations fonctionnelles pour le génome entier du pommier et disponible publiquement. Pour réaliser un enrichissement fonctionnel sur les données pommier, il faut alors utiliser l'organisme génétique modèle du végétal : *Arabidopsis thaliana*. La correspondance entre les gènes d'*Arabidopsis* et du pommier est disponible via l'identification des orthologues. Les orthologues sont des gènes de deux espèces différentes qui ont évolué depuis un ancêtre commun et sont donc séparés par un événement de spéciation. On considère donc qu'ils ont gardé la même fonction biologique. Ces relations d'orthologie entre *Arabidopsis* et le pommier s'appuient sur des comparaisons de séquences [Velasco *et al.*, 2010] et restent donc des prédictions. Avec les orthologues prédits, il est alors possible de faire une analyse fonctionnelle de données pommier.

Nous avons donc créé une annotation spécifique au pommier, en utilisant les orthologues d'*Arabidopsis*. Nous avons associé à chaque gène du pommier possédant un orthologue, les mêmes fonctions biologiques que l'orthologue d'*Arabidopsis*. Chaque gène du pommier qui ne possède pas d'orthologue se voit attribuer une fonction spécifique « inconnue », présente dans chacune des ontologies de la GO. Avec cette annotation, nous pouvons utiliser la totalité des transcrits d'intérêt dans l'analyse. La table 4.2 montre le nombre de transcrits du pommier qui ont un orthologue chez *Arabidopsis*. On voit ainsi que sur l'ensemble du génome du pommier (63 011 transcrits), 51 804 ont un orthologue avec *Arabidopsis* représentant 14 707 gènes d'*Arabidopsis* différents. Nous avons donc créé une annotation pour les 11 207 transcrits du pommier qui n'ont pas d'orthologue prédit avec *Arabidopsis* : cette annotation associe les gènes avec le terme “unknown biological process”.

TABLE 4.2 – Dénombrement de transcrits spécifiques à la fois pour le pommier et pour *Arabidopsis*.

Nombre de	Pommier	<i>Arabidopsis</i>
<i>Gènes annotés</i>	63 011	30 263
<i>Orthologues</i>	51 804	14 707
<i>Transcrits d'intérêt sens</i>	931	698
<i>Transcrits d'intérêt anti-sens</i>	694	524
<i>Transcrits d'intérêt</i>	1 625	1 222

L'analyse fonctionnelle différentielle s'effectue grâce à l'utilisation de BiNGO. En effet dans notre situation, nous avons besoin de travailler avec des annotations spécifiques au pommier et spécifiques aux données sens et anti-sens. L'utilisation d'outils en ligne qui reposent sur une ontologie prédéfinie dans

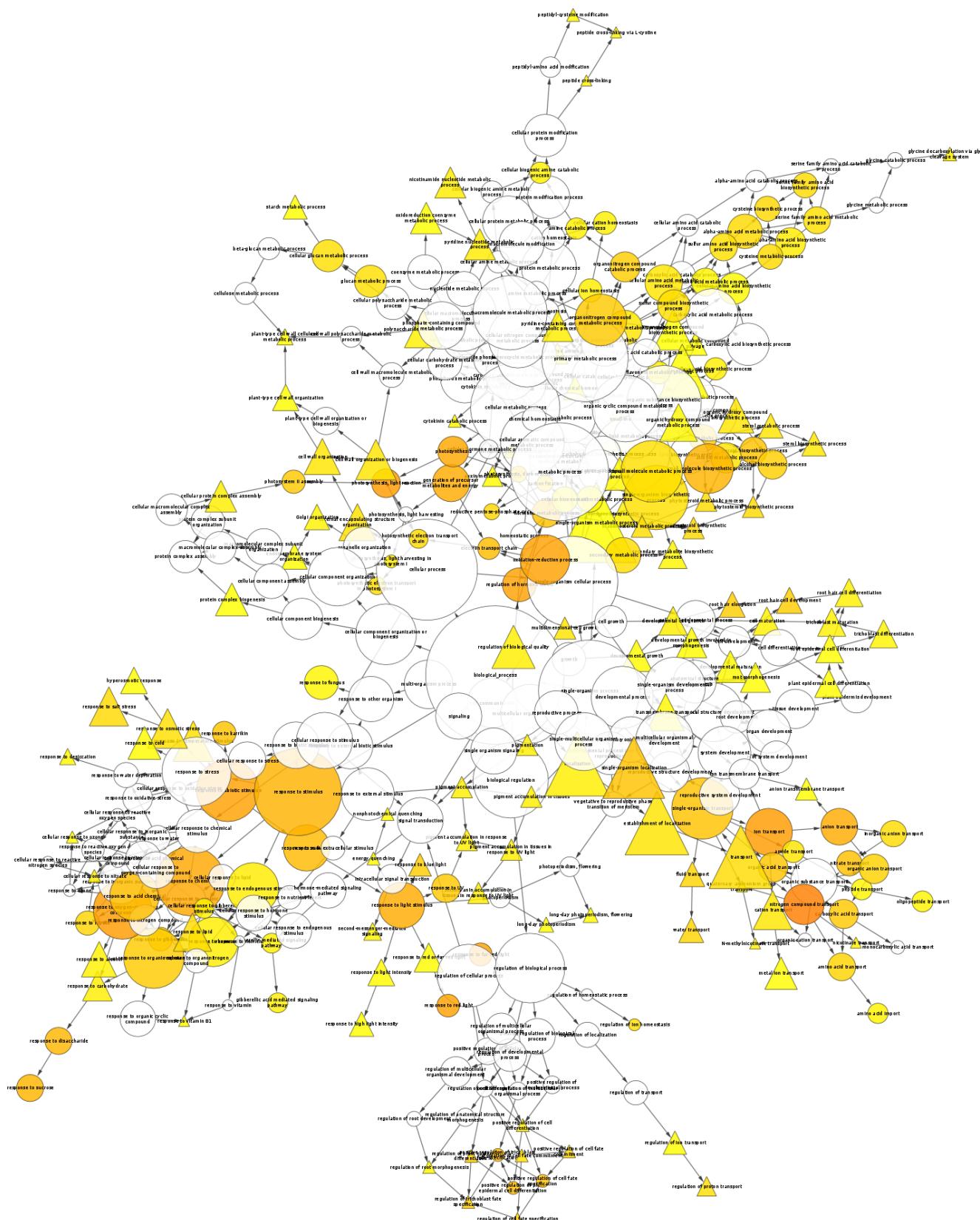


FIGURE 4.5 – Résultat graphique de l'analyse fonctionnelle différentielle H/60DAH. C'est la représentation graphique de l'analyse fonctionnelle des données sens et anti-sens pour laquelle les termes révélés par les anti-sens ont été mis en valeur grâce à une forme triangulaire.

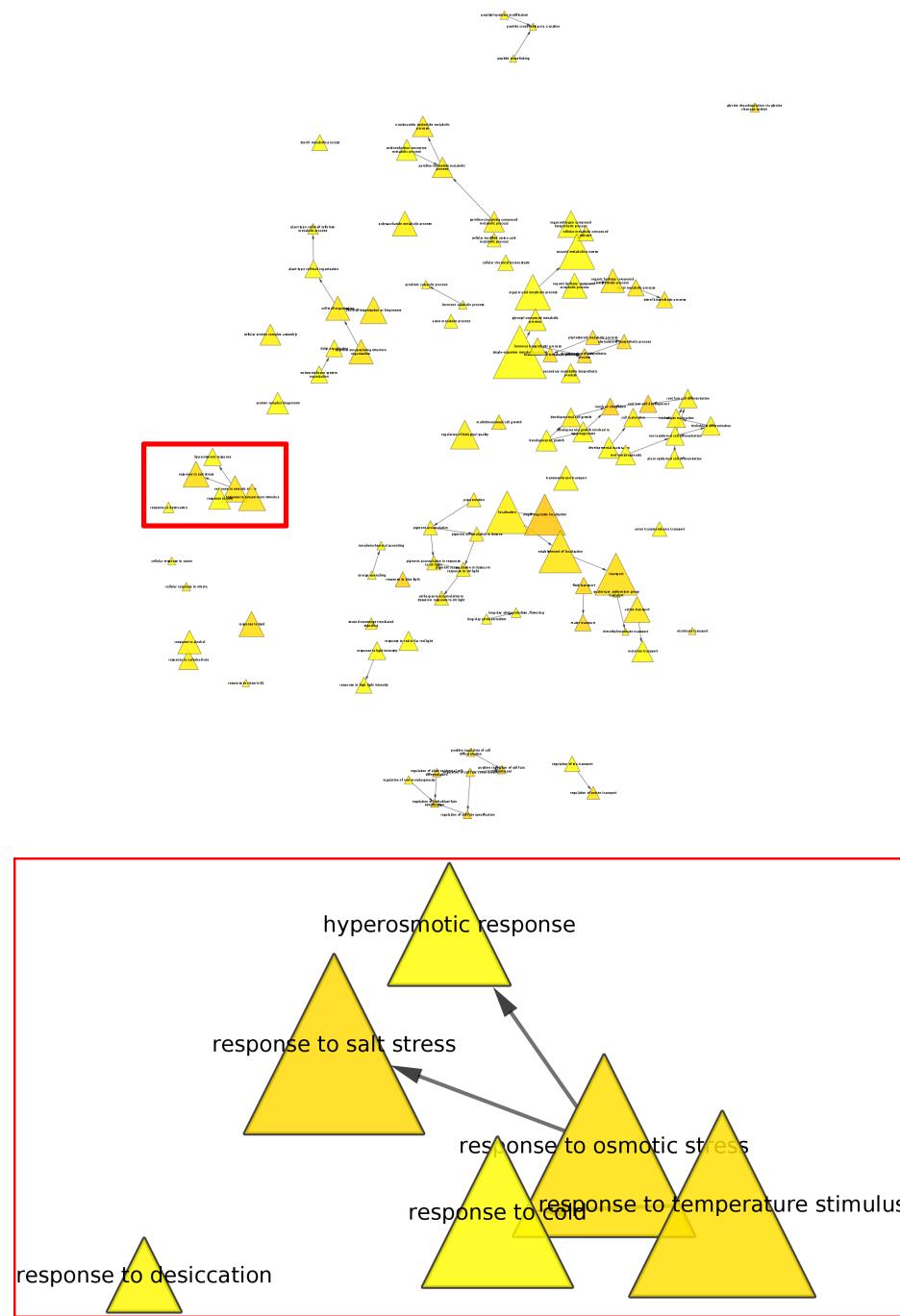


FIGURE 4.6 – Mise en valeur du résultat de l'analyse fonctionnelle différentielle. La partie haute est une représentation graphique de l'analyse fonctionnelle des données sens et anti-sens dans laquelle seuls les termes révélés par les anti-sens sont présents. La partie basse de l'image est le zoom du cadre rouge présent sur la partie haute. Ce zoom fait apparaître l'emplacement des termes “response to cold” et “hyperosmotic response”. La taille d'un nœud est proportionnelle au nombre de gènes rattachés à la catégorie GO. La couleur d'un nœud varie selon la sur-représentation de la catégorie GO (jaune : faible, orange : forte).

GO n'est donc pas possible. De plus la visualisation offerte par BiNGO est très intéressante pour l'exploitation des résultats. Nous considérons qu'une fonction est sur-représentées lors de l'enrichissement fonctionnel si sa p-valeur est inférieure à 0.05.

La figure 4.5 montre le résultat de l'analyse réalisée avec l'annotation du génome complet du pommier. Cette visualisation montre l'ensemble des termes significativement sur-représentés par l'ensemble des transcrits d'intérêt. La couleur des nœuds témoigne de cette sur-représentation variant du jaune –faiblement sur-représenté– au orange –fortement sur-représenté–, les nœuds blancs ne sont pas sur-représentés mais permettent de faire le lien hiérarchique entre les termes de l'ontologie. Les termes révélés par les anti-sens sont mis en valeur : les nœuds associés sont triangulaires dans notre visualisation alors que les autres sont circulaires.

La visualisation offerte par BiNGO permet de naviguer dans l'ontologie interrogée et d'identifier les catégories impactées. La figure 4.6 est la reprise de la figure 4.5 dans laquelle on ne représente plus les termes qui ne sont pas des termes révélés par les anti-sens. On observe alors que les termes révélés par les anti-sens sont répartis dans la hiérarchie et touchent donc tous types de catégories GO. Un zoom sur la figure 4.6 indique où se situent les termes “response to cold” et “hyperosmotic response”, qui seront rediscutés dans le chapitre 6 car ils sont relatifs à la condition expérimentale de la maturation du fruit conservé en chambre froide.

Un extrait des termes révélés par les anti-sens, obtenus à partir de l'annotation du pommier, sont listés dans la table 4.3. L'ensemble des 104 termes révélés par les anti-sens obtenus est listé dans la table 6.2 du chapitre 6. Une interprétation biologique plus complète de ces résultats sera fournie dans ce chapitre. Parmi les termes révélés par les anti-sens, on peut voir des termes liés au transport de lipides, à la paroi cellulaire, ainsi qu'aux réponses osmotiques et contre le froid. Ces fonctions biologiques, en lien avec la réponse au stress et la maturation du fruit, sont cohérentes avec le contexte expérimental dans lequel les fruits sont stockés dans des chambres froides pendant que la maturation continue. Sans les données anti-sens, ces termes liés à ces conditions ne sont pas dans le résultat d'un enrichissement fonctionnel.

TABLE 4.3 – Liste de termes révélés par les anti-sens. Pour chaque terme, nous indiquons la p-valeur associée (ce critère est utilisé pour trier la table) et le nombre de transcrits annotés par le terme. La table affiche les 12 premiers et 12 derniers termes révélés par les anti-sens selon la p-valeur.

GO catégorie	p-valeur	Nombre de transcrits
single-organism localization	2.5178e-04	285
response to blue light	5.5327e-04	25
root hair cell development	5.5327e-04	36
root hair elongation	5.6530e-04	35
regulation of trichoblast fate specification	6.2544e-04	4
regulation of plant epidermal cell differentiation	6.2544e-04	4
regulation of cell fate specification	6.2544e-04	4
water transport	1.0017e-03	32
fluid transport	1.0017e-03	32
positive regulation of cell fate commitment	1.0769e-03	4
brassinosteroid biosynthetic process	1.6036e-03	22
phytosteroid biosynthetic process	1.6036e-03	22
:	:	:
cellular metabolic compound salvage	4.0913e-02	26
anion transmembrane transport	4.1764e-02	23
hyperosmotic response	4.2619e-02	38
metal ion transport	4.2619e-02	63
response to high light intensity	4.2880e-02	31
root morphogenesis	4.3141e-02	52
plant epidermal cell differentiation	4.3141e-02	44
oxidoreduction coenzyme metabolic process	4.3698e-02	59
single-organism metabolic process	4.5237e-02	505
starch metabolic process	4.8270e-02	29
organonitrogen compound biosynthetic process	4.9633e-02	107
protein complex biogenesis	4.9633e-02	65

5

Analyse différentielle de réseaux

Nous présentons dans ce chapitre la méthode d'analyse différentielle de réseaux que nous proposons afin d'étudier l'impact de la transcription anti-sens sur les réseaux de gènes.

Comme on l'a rappelé dans le chapitre 2, l'analyse différentielle de réseaux est une approche développée depuis une dizaine d'années et qui permet, entre autres, de comparer des réseaux de gènes inférés à partir de données d'expression. Cette analyse permet de mettre en évidence des différences de co-expression, des différences d'interaction qui témoignent de régulations différentes lorsque les cellules sont dans des contextes différents.

Dans notre travail, nous intégrons dans nos analyses les transcrits anti-sens dont l'action a été jusqu'ici peu étudiée à large échelle. Afin d'explorer leurs rôles potentiels, notre méthode d'analyse différentielle compare un réseau inféré à partir de données sens uniquement et un réseau inféré à partir des données sens et anti-sens. L'inférence de réseaux est un problème difficile et les méthodes de reconstruction de réseaux obtiennent des résultats encore entachés de beaucoup d'erreurs. Lorsque l'on veut comparer deux réseaux inférés, les erreurs peuvent se cumuler. En effet, une interaction qui figure dans un réseau R_1 et ne se retrouve pas dans réseau R_2 peut être un faux positif de l'inférence de R_1 ou un faux négatif de l'inférence de R_2 . Pour limiter cet effet, nous restreignons notre comparaison aux interactions les plus fortes, rassemblées au sein de ce que nous appelons un cœur de réseau étendu. Cette comparaison nous permet d'identifier les transcrits sens qui voient leurs interactions fortement modifiées par la prise en compte des données anti-sens ; nous appelons ces gènes des *gènes AS-impactés* et nous définissons également les *motifs de changement* autour de ces gènes AS-impactés ; ces motifs sont formés d'un petit ensemble de gènes dont les inter-relations méritent une investigation plus poussée, par exemple à travers de nouvelles expérimentations biologiques. Lorsque les gènes AS-impactés sont reliés dans le cœur de réseau sens, nous proposons également d'étudier comment leurs connexions évoluent dans le réseau issu

des données sens et anti-sens.

La chapitre débute par la méthode d'inférence que nous proposons pour obtenir un cœur de réseau étendu. Ensuite nous définissons notre méthode d'analyse différentielle de réseaux, ainsi que les gènes AS-impactés et les motifs de changements qu'elle permet d'identifier. Nous présentons également l'algorithme qui nous permet d'étudier les reconfigurations de relations entre gènes AS-impactés. Nous donnons enfin en sections 5.2.2 et 5.3.3 les résultats de ces méthodes sur les données pommier.

5.1	Extended Core Network	97
5.1.1	Algorithme	97
5.1.2	Évaluation sur des données simulées	99
5.2	Analyse différentielle de réseaux	101
5.2.1	Gènes AS-impactés et motifs de changement	102
5.2.2	Résultats	107
5.3	Reconfiguration des interactions entre gènes AS-impactés	110
5.3.1	Problème de l'arbre de Steiner minimal	111
5.3.2	Méthode	113
5.3.3	Résultats	115

5.1 Extended Core Network

Comme nous l'avons rappelé dans le chapitre 2, de nombreuses méthodes d'inférence de réseaux de gènes à partir de données transcriptomiques ont été proposées. Nous avons présenté la méthode Conservative Causal Core Network (C3NET) [Altay et Emmert-Streib, 2010] qui propose de n'inférer qu'un « cœur de réseau ». A partir de la matrice de toutes les informations mutuelles entre paires de gènes, un cœur de réseau est obtenu en ne sélectionnant pour chaque gène qu'une unique interaction, définie par la valeur maximale d'information mutuelle trouvée entre ce gène et chacun des autres.

Dans notre étude, notre but est de comparer deux cœurs de réseaux pour identifier les changements significatifs observés lorsque l'on intègre les données anti-sens. Dans ce cas, considérer uniquement l'interaction de valeur maximale pour chaque gène est trop restrictif, puisque plusieurs valeurs d'informations mutuelles peuvent être très proches de la valeur maximale. Nous proposons donc une méthode d'inférence de réseaux de gènes, inspirée de C3NET, nommée *Extended Core Network (ECN)*, où pour chaque gène « les meilleures » interactions seront placées dans le cœur de réseau. Afin de définir une meilleure interaction, nous utilisons un taux d'acceptation qui va indiquer quelles valeurs proches de la valeur d'information mutuelle maximale seront retenues.

5.1.1 Algorithme

Pour un ensemble de transcrits considérés, la méthode Extended Core Network utilise en entrée une matrice contenant les valeurs d'information mutuelle estimées pour chaque paire de gènes. De la même manière que C3NET ou ARACNE, un test de significativité est tout d'abord appliqué sur cette matrice pour ne retenir que les valeurs significatives. Comme expliqué dans le chapitre 2, une valeur est significative si elle est supérieure à un seuil I_0 fixé par permutations des données. Toute valeur non-significative est remplacée par 0 avant d'appliquer la méthode d'inférence [Butte et Kohane, 2000].

L'algorithme ECN se décompose en deux étapes : la première étape *Extended Core Network Best Interactions (ECN_BI)* recherche pour chaque gène ses meilleures interactions et la seconde partie crée à partir de là un cœur de réseau qui est un graphe non-orienté.

L'algorithme ECN_BI (algorithme 1) calcule la matrice d'adjacence pondérée d'un graphe orienté pour un ensemble de gènes (ou de transcrits) G donné ; ce graphe indique pour chaque gène quelles sont ses meilleures interactions. Dans cet algorithme, nous initialisons d'abord la matrice d'adjacence à une matrice nulle. Ensuite, pour chaque gène, nous déterminons ses meilleures interactions c'est-à-dire ses voisins dans le graphe résultant. Les voisins d'un gène g sont ceux avec qui le gène g partage une information mutuelle maximale ou proche du maximum. Nous utilisons un taux d'acceptation r qui est un nombre compris entre 0 et 1 afin de définir l'intervalle de valeurs proches du maximum qui définit les meilleures valeurs. Pour un gène g , ayant avec les autres gènes une information mutuelle maximale notée t , un taux d'acceptation r signifie que toutes les valeurs d'informations mutuelles comprises dans $[t(1 - r); t]$ sont considérées comme les meilleures. Un taux d'acceptation de 0 signifie que seule la

Algorithme 1 ECN_BI(G, M, r)

Entrée : G : un ensemble de gènes.

Entrée : M : la matrice d'information mutuelle.

$M[i, j]$ est l'information mutuelle entre les gènes $i, j \in G$.

Entrée : r : le taux d'acceptation.

Sortie : A : la matrice d'adjacence du réseau de gènes.

{Initialisation de la matrice d'adjacence}

pour chaque $i, j \in G$ **faire**

$A[i, j] \leftarrow 0$

fin pour

pour chaque $i \in G$ **faire**

 {Calcul du seuil d'information mutuelle pour le gène i }

$t_i \leftarrow \max_{j \in G \setminus \{i\}} M[i, j]$

$t_i \leftarrow t_i * (1 - r)$

 {Calcul de la matrice d'adjacence}

pour chaque $j \in G \setminus \{i\}$ **faire**

si $M[i, j] \neq 0$ **et** $M[i, j] \geq t_i$ **alors**

$A[i, j] \leftarrow M[i, j]$

fin si

fin pour

fin pour

retourner A

valeur maximale de l'information mutuelle est considérée. Dans ce cas, la méthode est presque identique à C3NET ; en effet, si la valeur maximale d'information mutuelle est partagée par plusieurs interactions, toutes ces interactions seront intégrées au graphe avec ECN_BI alors qu'une seule sera sélectionnée par C3NET. Un taux d'acceptation de 1 signifie que toutes les interactions ayant été filtrées comme significatives sont intégrées au réseau de gènes ; mais pour obtenir un cœur de réseau, on envisage de travailler avec des valeurs faibles pour r .

En déterminant pour chaque gène ses meilleures interactions, l'algorithme ECN_BI (algorithme 1) détermine un graphe orienté, dans lequel les arcs vont d'un gène vers ses meilleurs voisins. Pour représenter ce graphe des meilleures interactions, nous utiliserons la représentation $g \rightarrow g'$ qui signifie que pour g , l'interaction entre g et g' figure parmi les meilleures.

Néanmoins l'information mutuelle est une mesure symétrique : l'information mutuelle entre les gènes g et g' est identique à l'information mutuelle entre g' et g . C'est pourquoi, comme cela est fait dans C3NET, l'algorithme ECN (algorithme 2) transforme le graphe orienté des meilleures interactions en un cœur de réseau non-orienté. Nous verrons dans la section 5.2 que la méthode d'analyse différentielle que nous proposons repose sur l'identification des meilleures interactions pour chaque gène, et donc sur le graphe orienté fourni par ECN_BI, mais en tant que méthode d'inférence d'un cœur de réseau, ECN doit retourner un graphe non-orienté.

La figure 5.1 montre les étapes de l'algorithme ECN sur une même matrice d'information mutuelle

Algorithme 2 ECN(M, r)

Entrée : M : la matrice d’information mutuelle.

$M[i, j]$ est l’information mutuelle entre les gènes $i, j \in G$.

Entrée : r : le taux d’acceptation.

Sortie : A : la matrice d’adjacence du réseau de gènes.

{Calcul des meilleurs voisins pour chaque gène}

$A \leftarrow \text{ECN_BI}(M, r)$

{Symétrisation de la matrice d’adjacence}

pour chaque $i \in G$ **faire**

pour chaque $j \in G \setminus \{i\}$ **faire**

si $A[i, j] \neq 0$ **alors**

$A[j, i] \leftarrow A[i, j]$

fin si

fin pour

fin pour

retourner A

mais avec deux taux d’acceptation différents. Cette figure permet de voir l’application de l’algorithme ECN_BI puis de l’étape de symétrisation, et elle permet également de voir l’effet du taux d’acceptation sur l’inférence des interactions. On remarque que la valeur d’information mutuelle 0.79 entre g_4 et g_3 est très proche de la valeur d’information mutuelle maximale 0.80 pour g_4 . Avec un taux d’acceptation de 0, seule la valeur maximale est sélectionnée pour former une interaction, on n’obtient ainsi que l’interaction entre g_4 et g_3 . Avec un taux d’acceptation de 0.05, on infère également une interaction entre g_4 et les gènes qui partagent une information mutuelle supérieure au seuil, c’est-à-dire à $0.80 * (1 - 0.05) = 0.76$. On infère ainsi une interaction entre g_4 et g_3 avec la méthode ECN et un taux d’acceptation de 0.05.

5.1.2 Évaluation sur des données simulées

Nous comparons ici les cœurs de réseaux obtenus par notre méthode Extended Core Network, pour différentes valeurs du taux d’acceptation r , avec ceux calculés par C3NET sur des données simulées. Pour simuler les données, nous nous basons sur des réseaux biologiques de référence et disponibles dans la littérature. À partir du réseau de référence, un sous-réseau de taille voulue est extrait ; ensuite une activité de ce sous-réseau est simulée pour produire les données d’expression artificielles ; la méthode d’inférence est alors appliquée sur ces données simulées et produit un réseau de gènes qui peut être comparé au sous-réseau original, ce qui correspond à une situation classique d’évaluation en apprentissage supervisé.

Nous avons utilisé SynTReN [Bulcke *et al.*, 2006] pour la sélection d’un sous-réseau biologique et la génération de données simulées ; nous avons travaillé avec un sous-réseau extrait de *E. coli* et un sous-réseau extrait de *S. cerevisiae*. Les méthodes d’inférence ont été testées dans un ensemble de $S = 500$ simulations. Une simulation $k \in [1..S]$ est lancée sur des données spécifiques X_k formées de n gènes et $j \in [\frac{p}{2}..p]$ échantillons aléatoirement sélectionnés à partir des données X générées par SynTReN. Cet ensemble de simulations a été réalisé 10 fois, de la sélection du sous-réseau biologique jusqu’aux 500

Matrice d'information mutuelle :

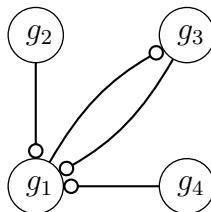
IM	g_1	g_2	g_3	g_4
g_1		0.70	1.10	0.80
g_2	0.70		0.60	0.50
g_3	1.10	0.60		0.79
g_4	0.80	0.50	0.79	

ECN avec un taux d'acceptation de 0 :

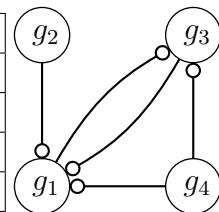
ECN avec un taux d'acceptation de 0.05 :

Extended Core Network Best Interactions :

A	g_1	g_2	g_3	g_4
g_1	0	0	1.10	0
g_2	0.70	0	0	0
g_3	1.10	0	0	0
g_4	0.80	0	0	0

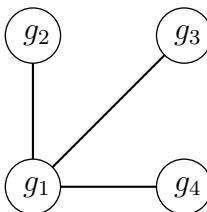


A	g_1	g_2	g_3	g_4
g_1	0	0	1.10	0
g_2	0.70	0	0	0
g_3	1.10	0	0	0
g_4	0.80	0	0.79	0



Symétrisation :

A	g_1	g_2	g_3	g_4
g_1	0	0.70	1.10	0.80
g_2	0.70	0	0	0
g_3	1.10	0	0	0
g_4	0.80	0	0	0



A	g_1	g_2	g_3	g_4
g_1	0	0.70	1.10	0.80
g_2	0.70	0	0	0
g_3	1.10	0	0	0.79
g_4	0.80	0	0.79	0

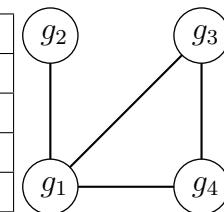


FIGURE 5.1 – Étapes de l'inférence de réseaux avec ECN et un taux d'acceptation de 0 (à gauche) et de 0.05 (à droite).

simulations.

Pour évaluer le taux d'erreur de chaque méthode d'inférence, nous comparons les interactions inférées avec les interactions présentes dans le réseau d'origine, ce qui donne une matrice de confusion. Une interaction est « vrai positif » si elle est présente à la fois dans le réseau inféré et dans le réseau original ; une interaction est « faux positif » si elle est présente uniquement dans le réseau inféré ; une interaction est « faux négatif » si elle est présente uniquement dans le réseau original ; une interaction est « vrai négatif » si elle n'est présente dans aucun réseau. Différentes courbes ou indices peuvent être utilisés pour évaluer la performance de l'inférence. La mesure de précision (équation 5.1) définit combien d'interactions sont de réelles interactions biologiques parmi les interactions inférées et le rappel (équation 5.2) définit la proportion des interactions réelles que l'on retrouve.

$$\text{précision} = \frac{\text{vrais positifs}}{\text{vrais positifs} + \text{faux positifs}} \quad (5.1) \quad \text{rappel} = \frac{\text{vrais positifs}}{\text{vrais positifs} + \text{faux négatifs}} \quad (5.2)$$

Ces deux indices sont souvent combinés pour obtenir la F -mesure qui est la moyenne harmonique de

la précision et du rappel, décrite par l'équation 5.3.

$$F = 2 \cdot \frac{\text{précision} \cdot \text{rappel}}{\text{précision} + \text{rappel}} \quad (5.3)$$

Nous comparons différents taux d'acceptation d'ECN avec C3NET sur les deux jeux de données issus d'*E. coli* et de *S. cerevisiae*. Nous testons ECN avec des taux allant de 0 à 1 avec un pas de 0.1, et de 0 à 0.2 avec un pas de 0.01.

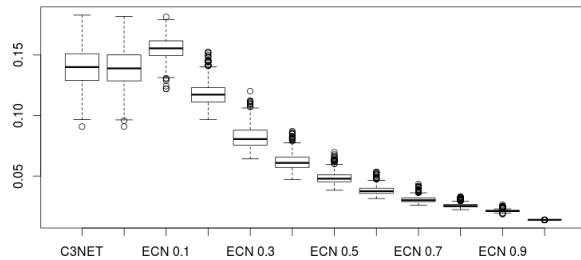
La figure 5.2 montre les box plots des F -mesures obtenus sur un des 10 ensembles de simulations. Tous les ensembles de simulations nous ont donné des résultats similaires.

La méthode ECN_0 correspond à ECN avec un taux d'acceptation de 0. Pour rappel, elle diffère de C3NET si plusieurs gènes g' partagent l'information mutuelle maximale avec le gène g . ECN a la même complexité que C3NET, c'est-à-dire en $\mathcal{O}(n^2)$, où n est le nombre de gènes. Nous pouvons voir cependant que la méthode ECN est meilleure que C3NET lorsque le taux d'acceptation est bas, à la fois pour *E. coli* (figures 5.2a et 5.2b) et pour *S. cerevisiae* (figures 5.2c et 5.2d). La F -mesure pour ECN dépend du taux d'acceptation. On peut observer que l'augmentation du taux d'acceptation a pour effet d'abord d'augmenter les performances avant de rapidement les dégrader. Comme expliqué plus tôt, un fort taux d'acceptation augmente le nombre de faux positifs, et on observe que cela impacte gravement la F -mesure. Cela s'observe sur les figures 5.2a et 5.2c, dès que le taux d'acceptation est respectivement supérieur à 0.1 et 0.2, la F -mesure chute jusqu'à devenir très faible. Néanmoins, lorsqu'on regarde les méthodes avec un taux d'acceptation entre 0 et 0.2 (figures 5.2b et 5.2d), on observe qu'un taux d'acceptation autour de 0.07 (ECN_0.07) est acceptable sur les données simulées. À partir de ces simulations, nous observons qu'un taux d'acceptation entre 0.05 et 0.1 est un bon compromis. Par la suite, nous avons donc décidé d'utiliser un taux d'acceptation de 0.05 avec les données pommier.

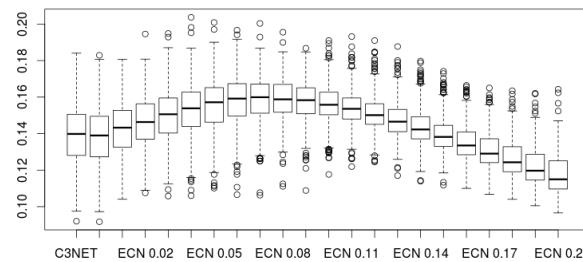
5.2 Analyse différentielle de réseaux

Depuis quelques années se développe un ensemble de travaux autour de l'analyse différentielle de réseaux et nous en avons présenté quelques-uns dans le chapitre 2. L'analyse différentielle de réseaux vise à mieux comprendre les régulations, et les dérégulations, propres à certains tissus ou certaines conditions expérimentales en comparant des réseaux d'interaction. Ces approches cherchent à aller au delà de l'observation de différences d'expression, pour identifier des différences de co-expression au sein d'un réseau de gènes.

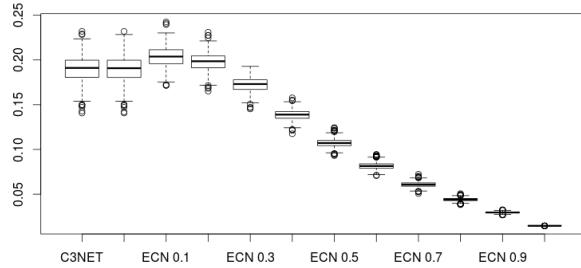
Afin d'étudier sur une large échelle l'impact de la transcription anti-sens, nous proposons une méthode d'analyse différentielle qui compare deux cœurs de réseaux de gènes inférés grâce à ECN. Il s'agit d'analyser les différences entre un réseau inféré à partir de données sens uniquement et un réseau inféré avec les données sens et anti-sens. Notre proposition diffère donc de l'analyse différentielle de réseaux pratiquée habituellement et qui compare des réseaux faisant intervenir les mêmes acteurs dans des conditions différentes. Ici nous proposons d'analyser l'intégration de nouveaux acteurs dans une condi-



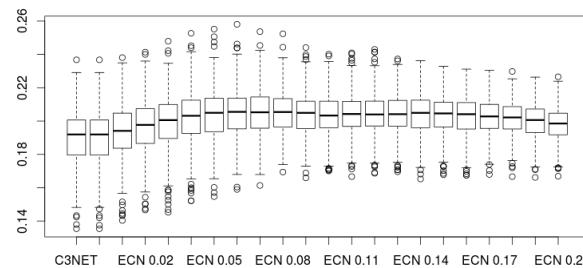
(a) *E. coli*, taux d'acceptation de ECN_0 jusqu'à ECN_1.



(b) *E. coli*, taux d'acceptation de ECN_0 jusqu'à ECN_0.2.



(c) *S. cerevisiae*, taux d'acceptation de ECN_0 jusqu'à ECN_1.



(d) *S. cerevisiae*, taux d'acceptation de ECN_0 jusqu'à ECN_0.2.

FIGURE 5.2 – Box plots des *F*-mesures pour C3NET et ECN avec différents taux d'acceptation. Le nombre suivant ECN indique le taux d'acceptation. La méthode C3NET est le premier box plot à gauche, suivi par les méthodes ECN triées en commençant par ECN_0. Les box plots sont obtenus à partir de 500 simulations sur deux ensemble de données : *E. coli* (a et b) et *S. cerevisiae* (c et d). Les taux d'acceptation de 0 jusqu'à 1 avec un pas de 0.1 sont testés (a et c) ainsi que de 0 jusqu'à 0.2 avec un pas de 0.01 (b et d).

tion expérimentale donnée. Le schéma de traitement que nous proposons est illustré dans la figure 5.3 (p. 103).

Pour une condition expérimentale donnée, nous considérons un ensemble de transcrits d'intérêt, des transcrits sens et des transcrits anti-sens. Notre analyse différentielle de réseaux compare un cœur de réseau inféré à partir de données sens, que l'on notera S_BI, avec un cœur de réseau inféré à partir de données sens et anti-sens, que l'on notera SAS_BI. Afin d'identifier quelles interactions maximales sont impactées par l'intégration des données anti-sens dans l'inférence du réseau, nous utilisons le réseau orienté obtenu grâce à l'algorithme ECN_BI, qui retient pour chaque gène l'ensemble de ses meilleures interactions.

Cette analyse nous amène à définir les notions de *gènes AS-impactés* et de *motifs de changement* révélés par la comparaison des deux réseaux.

5.2.1 Gènes AS-impactés et motifs de changement

Considérons un ensemble de transcrits sens G_S et un ensemble de transcrits anti-sens G_{AS} . Nous inférons un cœur de réseau orienté S_BI grâce à la méthode ECN_BI à partir de l'ensemble des transcrits G_S .

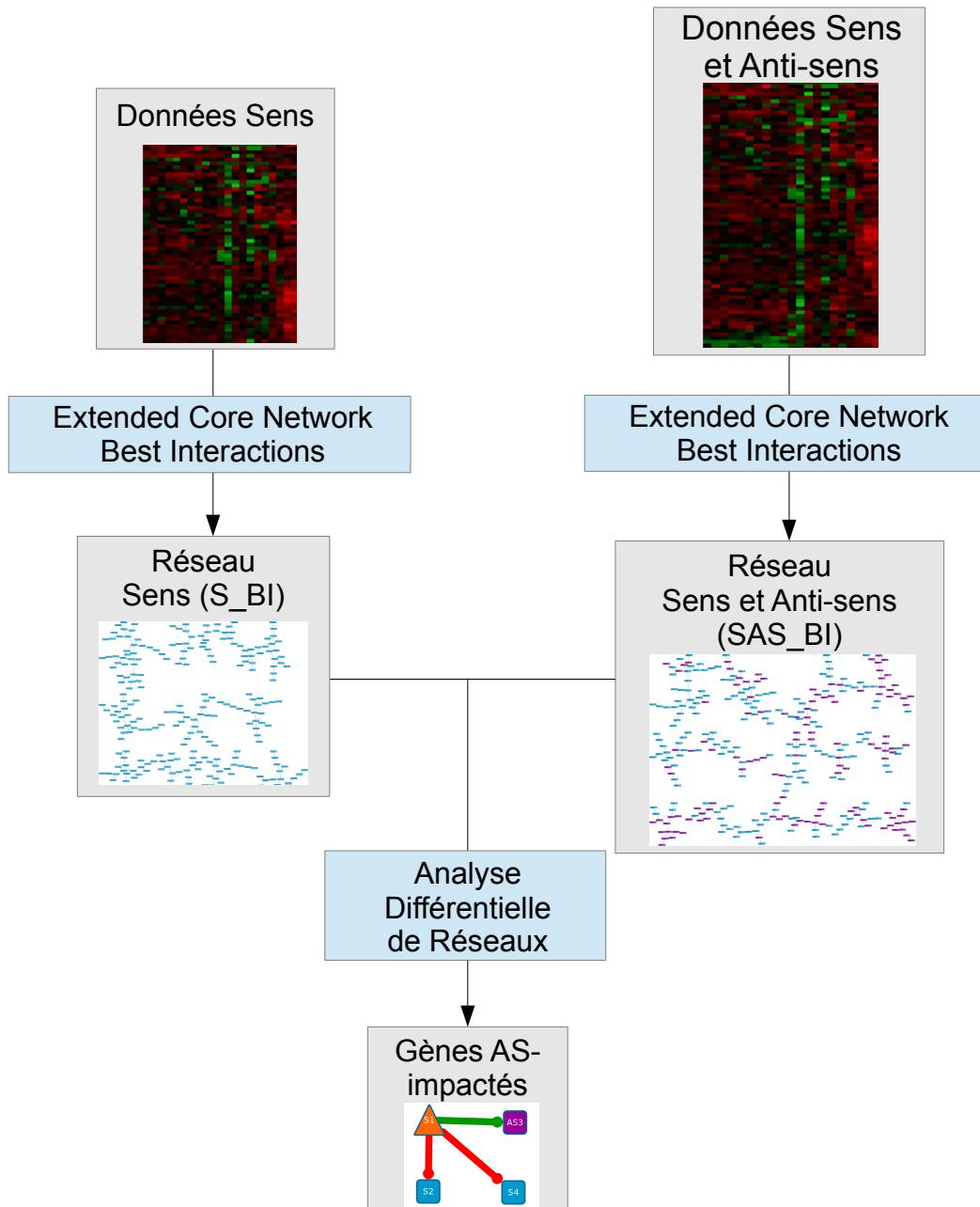


FIGURE 5.3 – Méthodologie de l’analyse différentielle de réseaux. Pour identifier des motifs de changement, nous réalisons une analyse différentielle de réseaux à partir de réseaux inférés avec Extended Core Network Best Interactions sur des données sens d’un côté, et des données sens et anti-sens de l’autre. Nous identifions enfin les changements d’interactions entre les différents réseaux.

Nous faisons de même à partir de l'ensemble des transcrits $G_S \cup G_{SAS}$ afin d'inférer le cœur de réseau orienté SAS_BI. Un *gène AS-impacté* (définition 5.1) est un gène sens qui interagit avec un ou plusieurs autres acteurs sens dans le cœur de réseau S_BI, mais ces interactions ne figurent pas dans le cœur de réseau SAS_BI. Ce changement dans les interactions se produit parce que dans le réseau SAS_BI, les meilleures interactions, c'est-à-dire les plus grandes valeurs d'information mutuelle, pour le gène AS-impacté concernent des acteurs anti-sens. Bien sûr, les valeurs d'information mutuelle entre le gène AS-impacté et ses voisins sens sont toujours identiques mais pour la constitution du cœur de réseau SAS_BI, elles ne figurent pas parmi les plus fortes valeurs, elles sont en quelque sorte « déclassées » par les valeurs d'information mutuelle que le gène AS-impacté partage avec un ou plusieurs anti-sens. Nous avons choisi ici une définition exigeante d'un gène AS-impacté : pour qu'un gène soit déclaré AS-impacté, il est nécessaire que toutes les meilleures interactions qu'il avait dans le cœur de réseau S_BI ne soient plus présentes dans le cœur de réseau de SAS_BI ; ainsi nous limitons notre analyse aux changements très significatifs provoqués par l'intégration des données anti-sens dans l'inférence de réseau.

Définition 5.1 (Gène AS-impacté). *Soient G_S un ensemble de transcrits sens et G_{AS} un ensemble de transcrits anti-sens. Soient $S_BI = (V_S, A_S)$ un réseau de gènes orienté inféré par ECN_BI à partir de G_S et $SAS_BI = (V_{SAS}, A_{SAS})$ un réseau de gènes orienté inféré par ECN_BI à partir de $G_S \cup G_{AS}$. Soient $N_S : V_S \rightarrow 2^{V_S}$ et $N_{SAS} : V_{SAS} \rightarrow 2^{V_{SAS}}$ les fonctions de voisinage respectivement de S_BI et de SAS_BI .*

Un gène $g \in G_S$ est un gène AS-impacté si, et seulement si :

$$N_S(g) \cap N_{SAS}(g) = \emptyset$$

La figure 5.4 illustre l'intégration des données anti-sens dans l'inférence de réseaux. La méthode ECN_BI est ici utilisée avec un taux d'acceptation de 0.05. La matrice d'information mutuelle des données sens est bien sûr incluse dans la matrice d'information mutuelle des données sens et anti-sens. Afin de différencier plus facilement les nœuds sens des nœuds anti-sens, nous représentons en bleu les nœuds sens et en violet les nœuds anti-sens. L'intégration des nouveaux acteurs anti-sens fait que les interactions inférées ne sont pas identiques. Bien sûr le réseau contient des interactions nouvelles pour les acteurs anti-sens ag_5 et ag_6 , mais on voit également que l'interaction de g_4 avec g_1 n'est pas représentée dans le cœur de réseau SAS_BI. L'information mutuelle entre le sens g_4 et l'anti-sens ag_5 est telle que la valeur seuil d'acceptation pour g_4 est passée de $0.70 \times 0.95 = 0.665$ dans le réseau S à $0.80 \times 0.95 = 0.76$. L'interaction de 0.70 n'est donc pas assez élevée pour être retenue dans le réseau SAS. g_4 est donc un gène AS-impacté.

L'identification des gènes AS-impactés est le premier résultat de notre analyse différentielle.

Ce résultat peut être exploité de différentes manières. Tout d'abord, notre implémentation fournit la liste des gènes AS-impactés, qui peut être directement utilisée par les experts des données.

La deuxième manière d'exploiter ce résultat est d'offrir une visualisation de ces gènes AS-impactés

Matrices d'information mutuelle :

IM_S	g_1	g_2	g_3	g_4
g_1		0.80	0.60	0.70
g_2	0.80		1.00	0.50
g_3	0.60	1.00		0.60
g_4	0.70	0.50	0.60	

IM_{SAS}	g_1	g_2	g_3	g_4	ag_5	ag_6
g_1		0.80	0.60	0.70	0.75	0.60
g_2	0.80		1.00	0.50	0.65	0.65
g_3	0.60	1.00		0.60	0.90	0.80
g_4	0.70	0.50	0.60		0.80	0.75
ag_5	0.75	0.65	0.90	0.80		0.85
ag_6	0.60	0.65	0.80	0.75	0.85	

Matrices d'adjacence :

S_BI	g_1	g_2	g_3	g_4
g_1	0	0.80	0	0
g_2	0	0	1.00	0
g_3	0	1.00	0	0
g_4	0.70	0	0	0

SAS_BI	g_1	g_2	g_3	g_4	ag_5	ag_6
g_1	0	0.80	0	0	0	0
g_2	0	0	1.00	0	0.65	0
g_3	0	1.00	0	0	0	0
g_4	0	0	0	0	0.80	0
ag_5	0	0	0.90	0	0	0
ag_6	0	0	0	0	0.85	0

Réseaux :

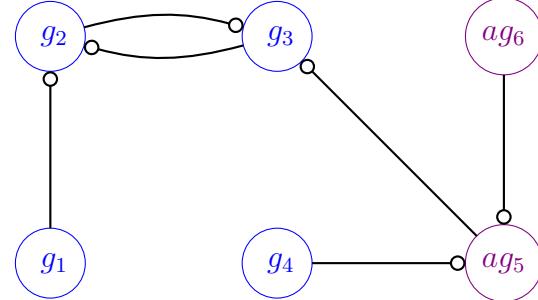
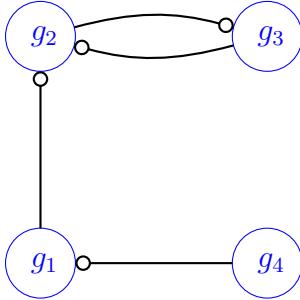


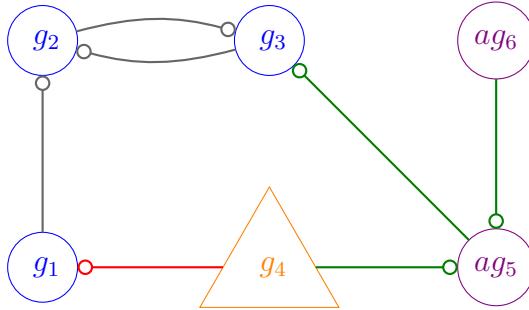
FIGURE 5.4 – Illustrations de l'intégration des données anti-sens dans l'inférence de réseau. Exemple de calcul des réseaux S_BI (à gauche) et SAS_BI (à droite) par la méthode ECN_BI avec un taux d'acceptation de 0.05, à partir des matrices d'information mutuelle.

dans un contexte de réseau. Les gènes AS-impactés sont des gènes de transcrits sens, ils se retrouvent donc tous dans le cœur de réseau S_BI, où nous proposons de les représenter. Cette visualisation est utile car elle permet d'observer si les gènes AS-impactés s'organisent d'une façon particulière ; on peut observer s'ils sont répartis dans les différentes zones du réseau, ou s'ils s'organisent plutôt en modules.

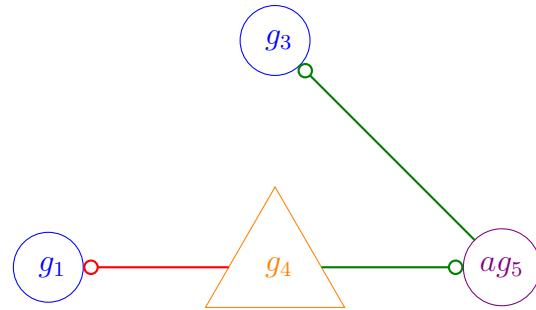
L'autre résultat de notre analyse différentielle est de mettre en évidence les interactions qui varient lorsque l'on prend en compte les données anti-sens. Un gène AS-impacté est défini dans un contexte de comparaison entre deux cœurs de réseau, nous souhaitons donc replacer le gène AS-impacté dans ce contexte pour voir les interactions qui ont permis de le définir. Afin de visualiser ces modifications d'interactions entre le réseau S_BI et le réseau SAS_BI, nous construisons un *graphe de changements* qui est une union des deux réseaux. Comme l'ensemble des nœuds du réseau S_BI est inclus dans l'ensemble des nœuds du réseau SAS_BI, le graphe de changements est le réseau SAS_BI auquel on ajoute les arcs

du réseau S_BI. Dans le graphe de changements, les arcs sont identifiés par des couleurs différentes qui indiquent à quel réseau ils appartiennent. La figure 5.5a montre un court exemple de graphe de changements. Un code couleur permet d'identifier rapidement la nature d'une interaction : rouge si elle est présente uniquement dans le réseau S_BI, verte si elle est présente uniquement dans le réseau SAS_BI ou grise si elle figure dans les deux réseaux. Nous gardons la même visualisation des noeuds sens et anti-sens mais nous ajoutons également une identification propre aux gènes AS-impactés : désormais ils sont dessinés par des triangles oranges.

La figure 5.5a est un exemple avec seulement six acteurs, la figure 5.6 (p. 108) montre le graphe de changement pour les données pommier. Le nombre d'interactions ainsi représentées dans le graphe de changements rend la lecture du graphique compliquée. De plus, les deux réseaux étant mélangés, il ne faut pas prendre en compte le placement des noeuds qui est biaisé par les interactions qui ne sont représentées que dans l'un ou l'autre des réseaux. L'intérêt de ce graphe est d'observer la modification des interactions.



(a) Graphes de changements.



(b) Motif de changement.

FIGURE 5.5 – Illustrations d'un *graphe de changements* et d'un *motif de changement* dans des réseaux inférés par ECN. Un nœud sens est représenté en bleu, un nœud anti-sens est représenté en violet. Le nœud triangulaire orange est un *gène AS-impacté*. Un lien rouge est un lien présent uniquement dans le réseau S. Un lien vert est un lien présent uniquement dans le réseau SAS.

L'identification des gènes AS-impactés nous permet d'observer localement les modifications apportées par l'intégration de données anti-sens. Avec le graphe de changements nous replaçons le gène AS-impacté dans son contexte : on observe les modifications d'interactions. Pour analyser biologiquement cet impact, nous regardons autour de cet AS-impacté quels sont les acteurs biologiques concernés par ces modifications. Nous extrayons ainsi des informations locales contenues dans le graphe de changements. Nous définissons les *motifs de changement* (définition 5.2) pour extraire ces informations.

Un motif de changement est composé de quatre types de gènes : le gène AS-impacté et les anti-sens qui impactent ses interactions forment le noyau du motif, et on y ajoute leurs voisins directs respectifs. On obtient ainsi (1) le gène AS-impacté, (2) les anti-sens impactants, (3) les voisins du gène AS-impacté dans le réseau S_BI et (4) les voisins des anti-sens impactants dans le réseau SAS_BI.

Définition 5.2 (Motif de changement). Soient G_S un ensemble de transcrits sens et G_{AS} un ensemble de transcrits anti-sens. Soient $S_BI = (V_S, A_S)$ un réseau de gènes orienté inféré par ECN_BI à partir de

G_S et $SAS_BI = (V_{SAS}, A_{SAS})$ un réseau de gènes orienté inféré par ECN_BI à partir de $G_S \cup G_{AS}$. Soient $N_S : V_S \rightarrow 2^{V_S}$ et $N_{SAS} : V_{SAS} \rightarrow 2^{V_{SAS}}$ les fonctions de voisinage respectivement de S_BI et de SAS_BI . Soit $g \in G_S$ un gène AS-impacté.

Le motif de changement autour de g est un graphe $M = (V_M, A_M)$ tel que :

- $V_M = \{g\} \cup N_S(g) \cup N_{SAS}(g) \cup \bigcup_{v \in N_{SAS}(g)} N_{SAS}(v)$
- $A_M = \{(v_1, v_2) \in A_S | v_1, v_2 \in V_M\} \cup \{(v_1, v_2) \in A_{SAS} | v_1, v_2 \in V_M\}$

Un motif de changement est un sous-graphe du graphe de changements. Les gènes contenus dans le motif de changement peuvent être identifiés en comparant les deux matrices d'adjacence des cœurs de réseaux S_BI et SAS_BI. La matrice sens permet d'identifier les voisins du gène AS-impacté. La matrice sens et anti-sens permet d'identifier les anti-sens impactants ainsi que leurs voisins. Cette identification peut également s'effectuer directement dans le graphe de changement en partant d'un gène AS-impacté puis, comme décrit dans la définition 5.2, en rajoutant ses voisins on obtient ainsi ses voisins sens avec les anti-sens impactants, enfin on rajoute les voisins de ces derniers pour obtenir le motif de changement.

Dans la figure 5.5a, il existe un seul motif : celui formé autour du gène AS-impacté g_4 (figure 5.5b) qui est impacté par ag_5 , le voisin de g_4 dans le réseau S est g_1 , le voisin de ag_5 dans le réseau SAS est g_3 . Le motif est donc ainsi constitué de (1) g_4 , (2) ag_5 , (3) g_1 et (4) g_3 .

Les motifs de changement permettent d'extraire des données un ensemble de gènes et de transcrits anti-sens qui méritent d'être étudiés. Les motifs de changement permettent ainsi de lire plus facilement un réseau de gènes et de mettre en évidence les effets d'une intégration de nouveaux acteurs.

5.2.2 Résultats

Nous présentons ici les résultats que nous avons obtenus avec notre méthode d'inférence Extended Core Network et l'analyse différentielle de réseau sur les données pommier. Une interprétation de ces résultats est fournie dans le chapitre 6 « Discussion biologique ».

Nous avons inféré des réseaux S et SAS avec ECN à partir des données pommier constituées de 931 transcrits sens et 624 transcrits anti-sens totalisant 1 625 transcrits.

Avec ECN et un taux d'acceptation de 0.05 on obtient un cœur de réseau S contenant 931 noeuds et 1 346 interactions et un cœur de réseau SAS contenant 1 625 noeuds et 2 377 interactions pour l'expérience 60DAH. Le réseau S de 60DAH est représenté dans la figure 5.7. On observe que le cœur de réseau est très connexe puisque composé de seulement 39 composantes connexes, dont une composante contient 832 gènes.

Nous inférons les réseaux S_BI et SAS_BI pour l'expérience 60DAH avec ECN_BI et un taux d'acceptation de 0.05 afin de rechercher des motifs de changement.

L'analyse différentielle des deux cœurs de réseaux nous donne 308 gènes AS-impactés définissant 308 motifs de changement. Ce premier résultat est à souligner puisque cela signifie qu'environ 30% des

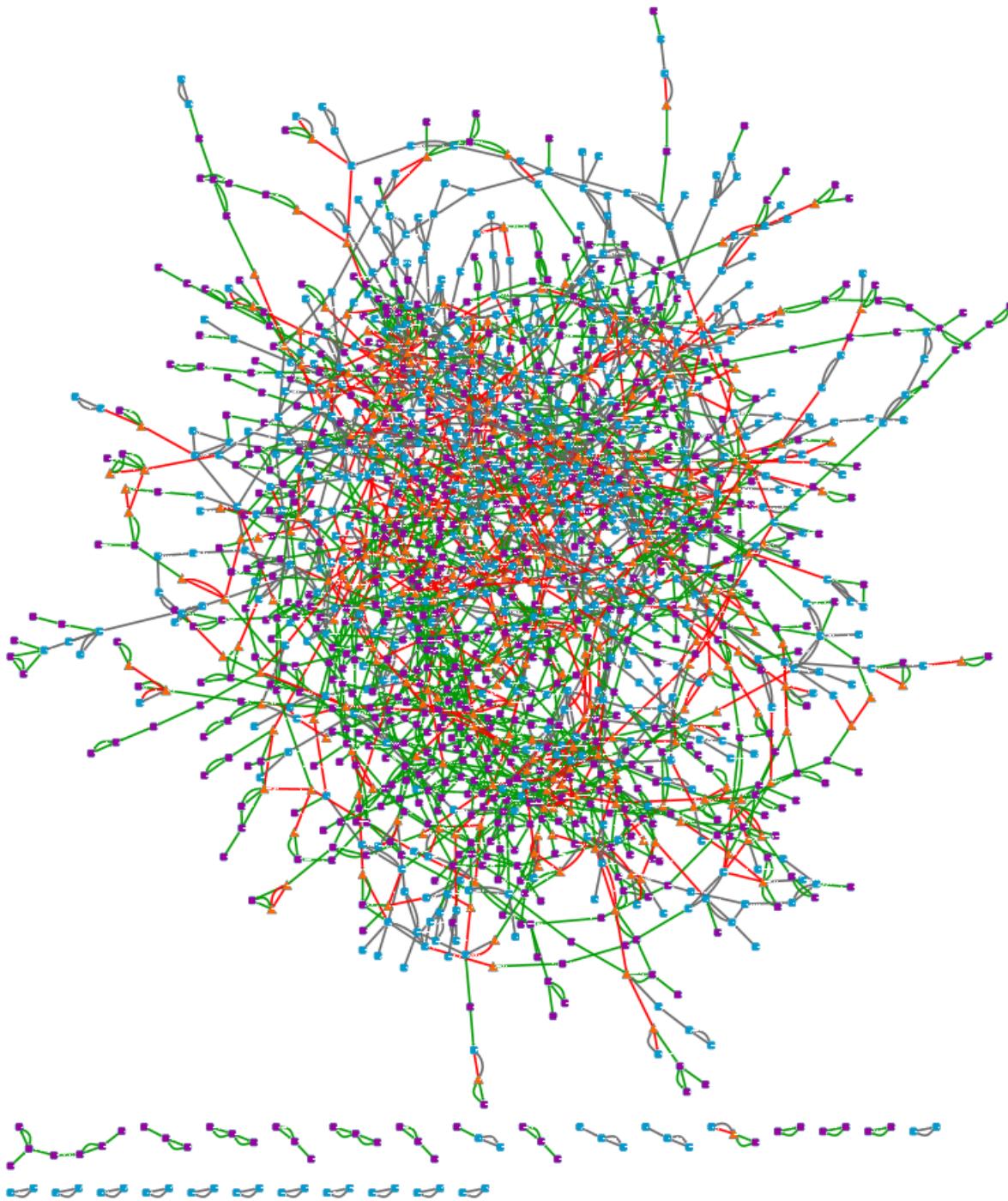


FIGURE 5.6 – Graphe de changements entre les réseaux de gènes S et SAS inférés avec ECN de l’expérience 60DAH. Un nœud bleu représente un gène sens ; un nœud violet représente un anti-sens ; un nœud triangulaire orange représente un gène AS-impacté. Un arc rouge représente un arc présent uniquement dans le réseau S ; un arc vert représente un arc présent uniquement dans le réseau SAS ; un arc gris représente un arc présent à la fois dans les réseaux S et SAS.

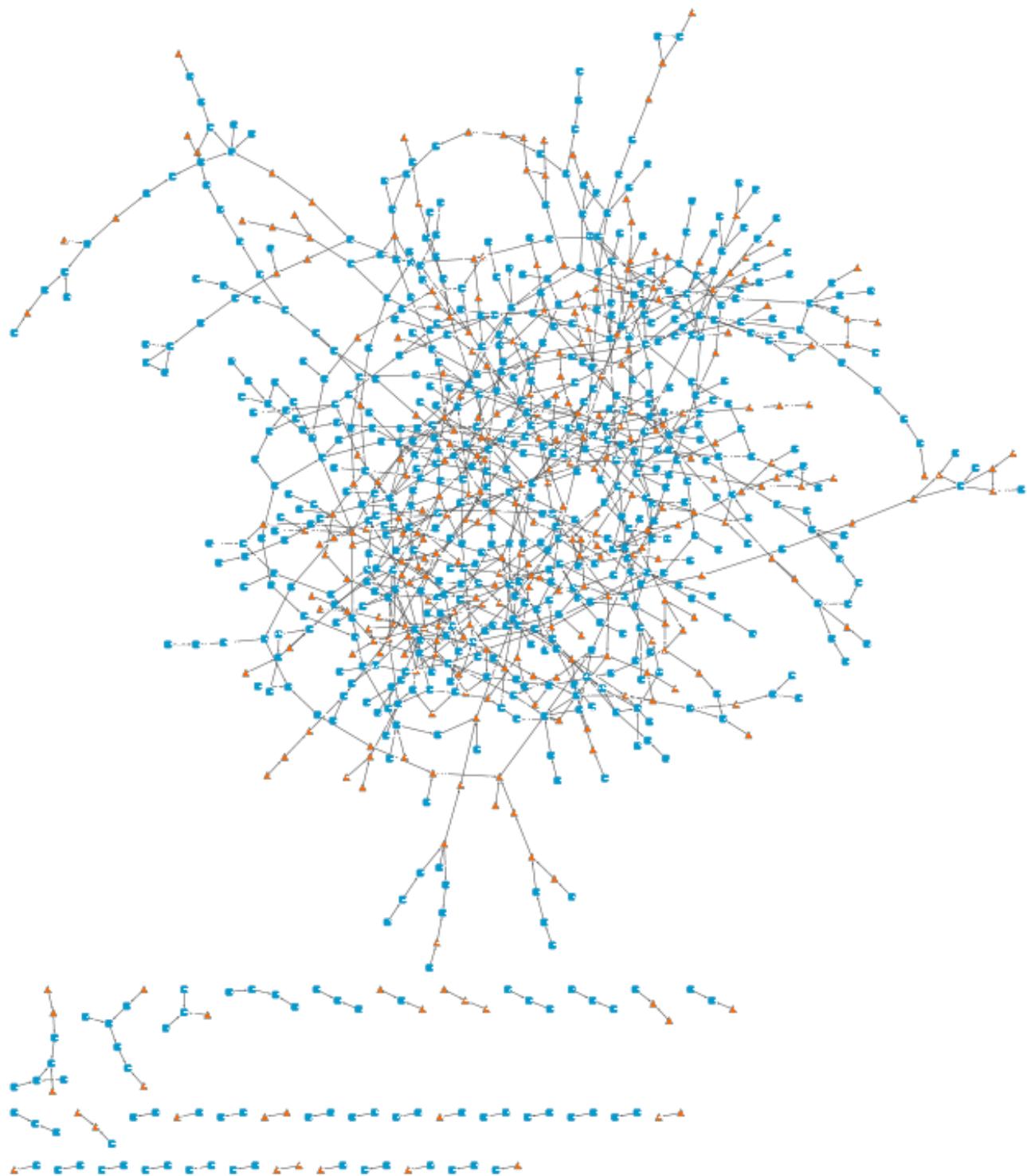


FIGURE 5.7 – Extended Core Network avec un taux d'acceptation de 0.05 pour les données sens de 60DAH. Les nœuds oranges triangulaires représentent les gènes AS-impactés.

931 gènes sens sont impactés par l'intégration des données anti-sens dans l'étude des réseaux. Cet impact est d'autant plus notable que ECN_BI infère des coeurs de réseau, ce sont donc les interactions principales qui sont impactées par l'intégration des données anti-sens.

Les motifs de changement permettent de mettre en lumière les gènes impactés mais également le voisinage de ces gènes. Une analyse fonctionnelle de ces gènes permet ensuite d'interpréter ces modifications. L'analyse de ces motifs de changement est disponible dans le chapitre « Discussion biologique ».

Après avoir identifié les gènes AS-impactés, grâce à ECN_BI, nous visualisons ces gènes dans le cœur de réseau inféré grâce à ECN afin de pouvoir les situer. Comme on peut le voir dans le réseau S de la figure 5.7, les gènes AS-impactés sont globalement répartis dans l'ensemble du cœur de réseau, ce qui signifie que différentes fonctions biologiques sont sans doute concernées. Mais on peut également remarquer que certains gènes AS-impactés sont connectés entre eux. Sur la figure 5.7, on observe facilement, dans une zone peu dense du graphe, une dizaine de gènes AS-impactés (triangles oranges) qui sont reliés. Sous Cytoscape, la visualisation interactive permet bien sûr une exploration plus poussée du graphique.

Nous savons que les interactions entre deux gènes AS-impactés ne seront pas représentées dans le réseau SAS, puisque, par définition, ces interactions ne seront plus parmi les meilleures interactions ni de l'un ni de l'autre. Nous souhaitons donc étudier la manière dont ces gènes interagissent dans le réseau SAS.

5.3 Reconfiguration des interactions entre gènes AS-impactés

Les motifs de changement fournissent une information locale sur les modifications des interactionsvenues après l'intégration des données anti-sens. Nous souhaitons également avoir une vue plus globale sur cet impact. Un gène AS-impacté est un gène qui possède une ou plusieurs interactions avec d'autres gènes sens dans le réseau S_BI, mais toutes ces interactions sont « déclassées » et ne sont pas assez significatives pour être représentées dans le réseau SAS_BI. Il nous paraît donc intéressant d'étudier la reconfiguration des interactions autour du gène AS-impacté, et plus précisément, il s'agit de voir si une interaction entre deux gènes AS-impactés dans le réseau S_BI, qui n'est plus présente dans SAS_BI, se retrouve mais de manière indirecte.

Nous souhaitons ainsi savoir si une interaction directe représentée dans le cœur de réseau S se retrouve de manière indirecte dans le réseau SAS. Dans la figure 5.4 (p. 105) on remarque ainsi que l'interaction de g_4 avec g_1 n'est plus représentée directement dans le réseau SAS, mais qu'on peut retrouver un lien entre ces gènes en utilisant le chemin : $g_4 - ag_5 - g_3 - g_2 - g_1$.

La figure 5.7 montre le réseau S obtenu pour l'expérience 60DAH, on peut remarquer que les gènes AS-impactés sont répartis dans l'ensemble du réseau, mais que beaucoup sont connectés entre eux et forment ainsi une partie du réseau pour laquelle la transcription anti-sens a un fort impact. Si deux gènes AS-impactés sont connectés dans S, on sait que cette interaction ne sera pas présente dans SAS. C'est pourquoi l'analyse de la reconfiguration ne s'effectue pas sur l'ensemble des gènes AS-impactés mais se

concentre sur les gènes AS-impactés qui interagissent entre eux. Cette analyse se base sur l'utilisation du problème de l'arbre de Steiner que nous détaillons ci-après, avant de présenter notre approche.

5.3.1 Problème de l'arbre de Steiner minimal

Étant donné un graphe non-orienté $G = (V, A)$ avec un ensemble de nœuds V et un ensemble d'arêtes A , le problème classique de l'arbre de Steiner (définition 5.3) est de trouver, pour un ensemble de nœuds T appelés nœuds terminaux, un sous-graphe G' de G contenant T de telle manière qu'il existe un chemin entre chaque paire de nœuds de T et avec un nombre minimal d'arêtes. Les nœuds de $V - T$, nécessaires à l'arbre de Steiner G' pour obtenir un graphe connexe, sont appelés nœuds de Steiner. Lorsque les arêtes sont pondérées, le problème de l'arbre de Steiner minimal est de trouver un sous-graphe qui recouvre tous les nœuds terminaux avec une somme des poids minimale.

Définition 5.3 (Arbre de Steiner minimal). *Soit un graphe $G = (V, A)$ et un ensemble de nœuds T tel que $T \subseteq V$. Un arbre de Steiner $G' = (V', A')$ est un sous-graphe de G tel que :*

- $T \subseteq V' \subseteq V$,
- $A' \subseteq A$,
- *il existe un chemin entre chaque paire de T dans G' .*

G' est un arbre de Steiner minimal si $|A'|$ est minimal.

Ce problème d'optimisation combinatoire est connu pour être NP-complet [Karp, 1972], c'est pourquoi plusieurs méthodes heuristiques ont été proposées pour résoudre ce problème sur des graphes de grande taille.

Pour calculer l'arbre de Steiner minimal, nous utilisons une approche heuristique appelée l'approximation par les plus courts chemins [Sadeghi et Fröhlich, 2013]. La méthode construit un arbre de Steiner ST par étapes successives. La première étape sélectionne arbitrairement un nœud terminal comme étant le premier nœud de l'arbre de Steiner ST . Ensuite l'algorithme recherche le nœud terminal qui est le plus proche de tous les nœuds actuellement dans ST , puis ajoute ce plus court chemin à l'arbre de Steiner en construction ST . La recherche s'arrête lorsque tous les nœuds terminaux sont dans ST . Finalement, l'arbre résultant est élagué afin d'obtenir l'arbre couvrant minimal de ST et tous les nœuds de Steiner de degré 1 sont retirés. Le but de l'arbre de Steiner minimal étant d'obtenir un arbre du plus petit poids, la dernière étape permet de retirer des interactions non-nécessaires à la connexion des nœuds terminaux et qui augmentent le poids de l'arbre.

L'heuristique des plus courts chemins est un bon compromis entre la qualité de la solution et le temps de calcul sur des grands graphes tels que ceux utilisés en bio-informatique [Sadeghi et Fröhlich, 2013]. La complexité de cet algorithme de recherche d'un arbre de Steiner dans un graphe $G = (V, E)$ pour les nœuds terminaux T est en $\mathcal{O}(|T| \cdot (|V| + |E|))$, $|V| + |E|$ étant la complexité de la recherche du plus

court chemin. Le résultat de cette approximation dépend cependant beaucoup du choix du premier nœud terminal lors de l'initialisation : c'est lui qui va déterminer par la suite la topologie de l'arbre.

Le paquet R SteinerNet [Sadeghi et Fröhlich, 2013] implémente cette approximation avec trois autres méthodes d'approximation et une méthode exacte pour résoudre le problème de l'arbre de Steiner minimal. Ce paquet n'a pas été mis à jour depuis 2013 et ses dépendances ne sont plus supportées par une version récente de R, il a donc été supprimé du dépôt CRAN. Afin de pouvoir l'utiliser avec la version 3.2.0 de R, nous avons mis à jour le paquet SteinerNet. Cette version à jour est disponible publiquement à l'adresse <http://www.info.univ-angers.fr/~legeay/SteinerTree.r>.

Les trois autres méthodes d'approximation du paquet SteinerNet sont : l'approximation basée sur l'arbre couvrant minimal (KB), l'approximation aléatoire de tous les plus courts chemins (RSP) et tous les plus courts chemins (ASP) que nous décrivons brièvement ci-dessous.

L'approximation basée sur l'arbre couvrant minimal (KB) est similaire à l'heuristique des plus courts chemins. Alors que l'heuristique des plus courts chemins travaille avec un graphe qui petit à petit devient l'arbre minimal de Steiner, l'approximation KB travaille avec un ensemble de sous-graphes qui, par fusion de sous-graphes, deviendra l'arbre minimal de Steiner. L'approximation basée sur l'arbre couvrant minimal considère au début que chaque nœud terminal est un sous-graphe et va fusionner ces sous-graphes au fur et à mesure. Les plus courts chemins entre les nœuds des différents sous-graphes sont calculés et le plus petit des plus courts chemins permet de fusionner les deux sous-graphes. Lorsqu'il ne reste plus qu'un sous-graphe, l'arbre couvrant minimum de ce sous-graphe est calculé et les nœuds non terminaux de degré 1 sont retirés de l'arbre. Cet arbre est considéré comme l'arbre minimal de Steiner. Afin de rechercher les plus courts chemins, la complexité est en $\mathcal{O}(|V| + |E|)$ par graphe, sachant qu'il y a au plus $|T|$ graphes on obtient une complexité en $\mathcal{O}(|T| \cdot (|V| + |E|))$. Les graphes sont ensuite fusionnés puis l'opération est répétée jusqu'à obtenir l'arbre, il y aura $|T|$ recherches des plus courts chemins soit une complexité de l'algorithme KB en $\mathcal{O}(|T|^2 \cdot (|V| + |E|))$.

L'approximation aléatoire de tous les plus courts chemins (RSP) calcule tous les plus courts chemins entre les nœuds terminaux, ce qui forme son graphe initial. Un arbre couvrant minimal T est calculé sur ce graphe initial. Ensuite un nœud de Steiner est choisi aléatoirement et supprimé du graphe, sauf si cette suppression forme deux composantes connexes. Un arbre couvrant minimal est calculé sur ce graphe et s'il est plus petit que le premier arbre couvrant T , la suppression du nœud est conservée, sinon le nœud est remis. Le processus s'arrête lorsqu'il n'est plus possible de supprimer un nœud de Steiner. Le calcul d'un arbre couvrant minimal s'effectue en $\mathcal{O}(|V| \log |E|)$, et le calcul des plus courts chemins s'effectue en $\mathcal{O}(|T| \cdot (|V| + |E|))$ donc l'algorithme a une complexité en $\mathcal{O}(|V| \log |E| + |T| \cdot (|V| + |E|))$.

Enfin la méthode ASP est simplement une fusion de l'ensemble des plus courts chemins entre tous les nœuds terminaux. La complexité de cet algorithme est donc en $\mathcal{O}(|T| \cdot (|V| + |E|))$.

Le problème de l'arbre de Steiner peut être utilisé pour récolter des informations de grandes bases de données d'interactions moléculaires [Dittrich *et al.*, 2008]. Par exemple, si on étudie l'implication d'un ensemble de protéines dans l'interactome, il peut extraire des bases de données dédiées un arbre de Steiner avec ces protéines comme nœuds terminaux. Les auteurs de [Sadeghi et Fröhlich, 2013] ont utilisé

les arbres de Steiner dans des réseaux de protéine-protéine chez l'Humain. Des gènes ont été identifiés comme jouant un rôle dans le cancer du sein et les auteurs ont recherché des arbres de Steiner minimaux qui permettent de relier ces gènes dans le réseau protéine-protéine. Ces arbres ont permis d'identifier des protéines qui interagissent avec le produit des gènes. La recherche des arbres de Steiner dans le réseau protéine-protéine a ainsi permis d'identifier de nouveaux acteurs dans le cancer du sein.

5.3.2 Méthode

Nous souhaitons donc étudier l'impact de la transcription anti-sens sur un groupe de gènes AS-impactés. La figure 5.8 montre le processus d'analyse de la reconfiguration des interactions des gènes AS-impactés, décrit ci-après. À partir des données transcriptomiques sens et anti-sens, nous inférons grâce à ECN_BI deux cœurs de réseaux orientés S_BI et SAS_BI qui nous permettent d'identifier les gènes AS-impactés. Nous identifions ces gènes AS-impactés dans le cœur de réseau S inféré avec ECN. Afin d'identifier où se situent les gènes AS-impactés connectés entre eux, nous calculons le sous-graphe de S induit par l'ensemble des gènes AS-impactés. Le sous-graphe de S induit par l'ensemble des gènes AS-impacté est le sous-graphe de S qui ne contient que les nœuds représentant un gène AS-impacté, et les interactions qui relient ces derniers entre eux.

Nous définissons ainsi un *sous-graphe AS-impacté* (définition 5.4) comme étant une composante connexe du réseau induit dans S par les gènes AS-impactés.

Définition 5.4 (Sous-graphe AS-impacté). *Soit un cœur de réseau $S = (V_S, A_S)$ inféré à partir de ECN sur des données sens. Soit $G_{\text{impacté}} \subset V_S$ l'ensemble des gènes AS-impactés de S .*

Soit $S_{\text{impacté}} = (G_{\text{impacté}}, A_{\text{impacté}})$ le sous-graphe de S induit par $G_{\text{impacté}}$.

Un sous-graphe AS-impacté est une composante connexe de $S_{\text{impacté}}$.

Un sous-graphe AS-impacté est donc uniquement composé de gènes AS-impactés qui sont connectés dans le réseau S. Comme ces interactions présentes dans le sous-graphe ne seront plus représentées dans le réseau SAS, il est intéressant de connaître les relations existant entre ces nœuds dans le réseau SAS. Une façon de répondre à ce problème est d'utiliser la notion d'arbre de Steiner.

Pour chaque sous-graphe AS-impacté du réseau S, nous recherchons un arbre de Steiner dans le réseau SAS qui reconnecte les gènes AS-impactés. Si l'arbre peut être trouvé, cet arbre de Steiner montre comment les interactions présentes dans le réseau S entre les gènes AS-impactés sont reconfigurées dans le réseau SAS. Remarquons que puisque nous souhaitons regarder les interactions entre les gènes d'un sous-graphe AS-impacté, nous n'analyserons que les sous-graphe AS-impactés composés d'au moins trois gènes : les gènes AS-impactés non reliés à d'autres AS-impactés ne seront donc pas analysés.

L'information apportée par notre analyse peut être exploitée de différentes manières. D'une part, une visualisation de la reconfiguration des interactions peut aider à détecter des interactions intéressantes. D'autre part, une analyse fonctionnelle des gènes AS-impactés peut être réalisée pour rechercher des fonctions impactées par la transcription anti-sens.

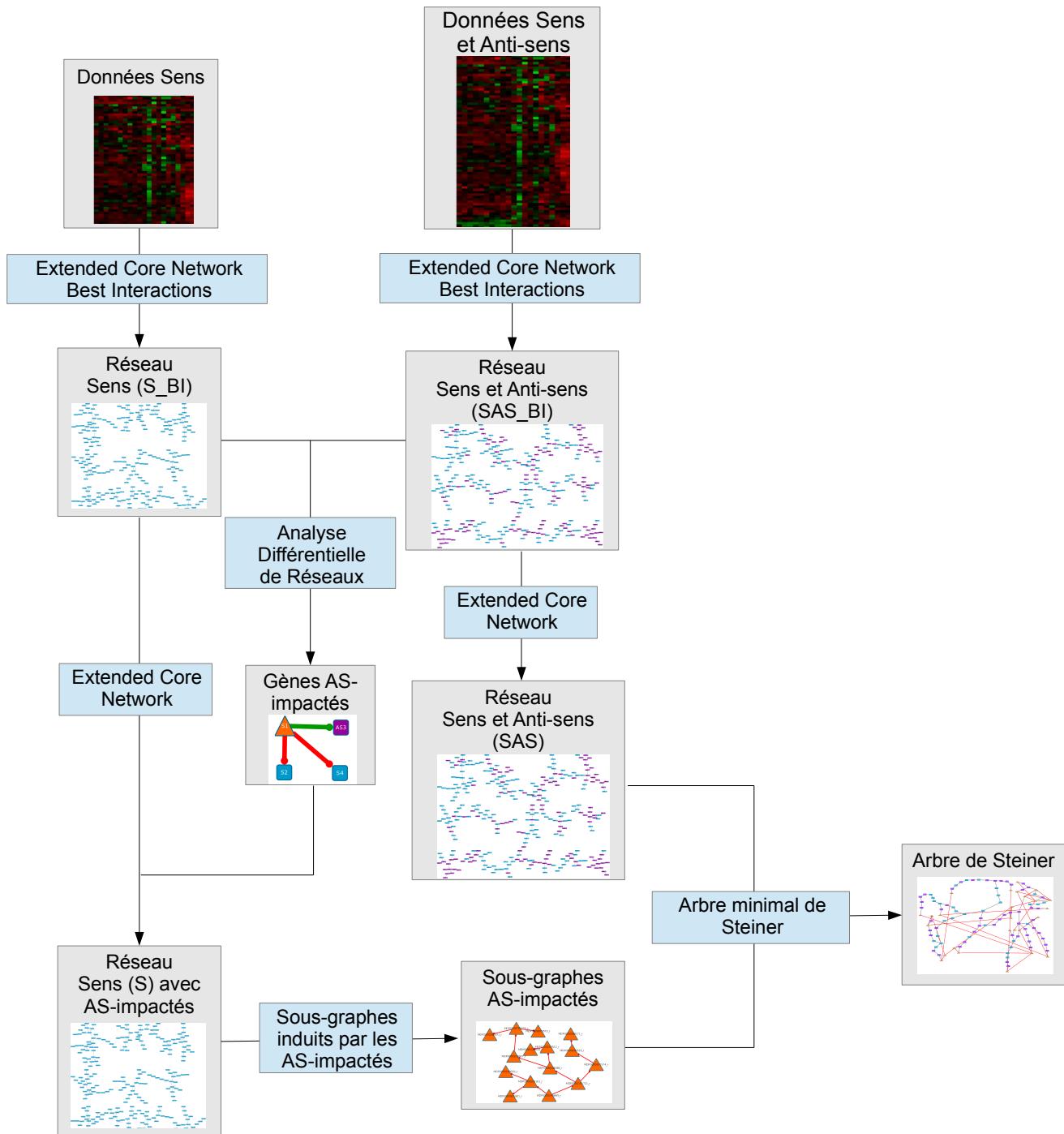


FIGURE 5.8 – Processus d’analyse de la reconfiguration des interactions des gènes AS-impactés.

Nous avons jusqu’ici utilisé les versions orientées des réseaux de gènes fournies par ECN_BI afin d’identifier où se situent les modifications d’interactions. Dans l’étude de la reconfiguration des interactions entre gènes AS-impactés, nous nous plaçons dans une vue plus globale des interactions. Pour la recherche des arbres de Steiner, nous utilisons donc le réseau SAS non-orienté fourni par ECN. L’information mutuelle utilisée pour inférer les réseaux de gènes étant symétrique, nous pouvons ainsi observer la reconfiguration des interactions telle qu’observée dans un réseau de gène classique.

Dans la recherche de l’arbre de Steiner minimal, nous utilisons l’approximation des plus courts chemins. L’algorithme des plus courts chemins s’effectue sur des graphes pondérés : soit toutes les interactions sont identiques et alors un poids de 1 est affecté à chacune, ou alors chaque interaction a un propre poids. Dans notre cas, nous pouvons envisager la recherche de l’arbre de Steiner minimal avec deux mesures de distances. Soit nous considérons que toutes les interactions du réseau SAS ont le même poids. Soit nous faisons intervenir une connaissance biologique en cherchant des chemins qui relient des gènes présentant une certaine similitude fonctionnelle. En effet, on peut, à partir des connaissances accumulées dans la Gene Ontology, définir une mesure de distance sémantique entre gènes. C’est ce qui est proposé dans le paquet R GOSemSim [Yu *et al.*, 2010] que nous avons utilisé. GOSemSim fournit la distance sémantique entre deux gènes : c’est une mesure qui indique la distance entre les termes de la GO associés aux deux gènes. Ainsi, plus deux gènes ont des catégories GO proches selon la hiérarchie de l’ontologie, plus ils auront une distance sémantique faible. Le paquet permet donc de calculer une matrice de distances sémantiques lorsqu’on lui fournit une liste de gènes. Cette matrice permet donc d’affecter des poids aux interactions et d’en tenir compte lors de la recherche de l’arbre de Steiner minimal.

5.3.3 Résultats

Dans le cœur de réseau S nous avons 931 noeuds sens et nous avons identifié 308 gènes AS-impactés. Nous dénombrons 142 sous-graphes AS-impactés pour l’expérience 60DAH et nous travaillons avec les 29 sous-graphes AS-impactés qui possèdent au moins trois gènes. Ces sous-graphes AS-impactés sont recherchés dans le réseau SAS. Il est possible que les gènes d’un sous-graphe AS-impacté ne se retrouvent pas dans une même composante connexe du réseau SAS, il peut ainsi y avoir plusieurs arbres de Steiner pour un sous-graphe ou aucun. On trouve 35 arbres de Steiner pour l’expérience 60DAH puisque six sous-graphes AS-impactés ont fourni deux arbres de Steiner chacun.

Lorsque les éléments d’un sous-graphe AS-impacté ne peuvent pas être reconnectés par un arbre de Steiner, cette information est également intéressante ; cela signifie que les interactions représentées dans le réseau S ne se retrouvent pas du tout dans le réseau SAS.

La recherche des arbres minimaux de Steiner a été effectuée en utilisant à la fois la distance classique et la distance sémantique. On n’observe pas de différence entre l’utilisation des deux distances. Cela peut s’expliquer de plusieurs manières. La première explication est que nous recherchons des arbres de Steiner dans un cœur de réseau, c’est-à-dire un graphe avec peu de liens, et donc s’il n’existe qu’un seul chemin pour relier deux gènes, ce chemin est forcément le plus court quelle que soit la distance utilisée. La

seconde explication est que les gènes voisins sont très similaires sémantiquement, ainsi lorsque plusieurs chemins sont possibles, la longueur de ces chemins sont très proches. En effet, nous avons remarqué que dans notre réseau, GOSemSim affecte des distances très proches aux différents liens entre un gène et ses différents voisins. Cette distance n'est donc pas assez discriminante pour distinguer les interactions et les chemins qui en résultent.

La figure 5.9 montre un arbre de Steiner pour l'expérience 60DAH. Dans cette figure, on y voit à la fois les interactions du réseau S et du réseau SAS. On observe alors que certains gènes AS-impactés qui étaient voisins directs se retrouvent maintenant à plus d'une dizaine de gènes de distance. L'arbre de Steiner permet d'identifier les acteurs qui permettent de « reformer » les interactions directes observées dans le réseau S.

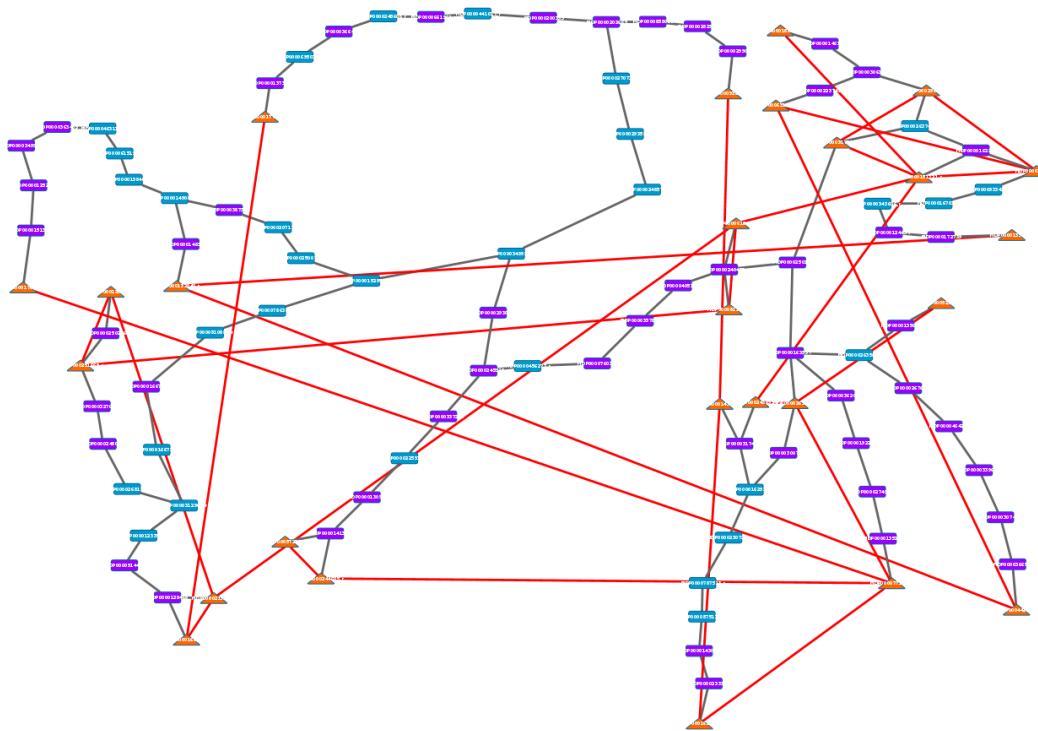
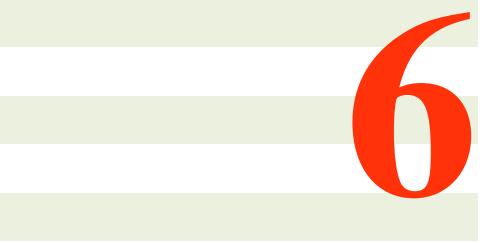


FIGURE 5.9 – Arbre de Steiner d'un sous-graphe AS-impacté de 60DAH. Les noeuds oranges sont les noeuds terminaux donc des gènes AS-impactés ; les noeuds bleus sont les noeuds sens de Steiner ; les noeuds violets sont les noeuds anti-sens de Steiner. Les liens gris sont les connexions du réseau SAS ; les liens rouges sont les connexions du réseau S.

Les sous-graphes AS-impactés permettent ainsi de mettre en évidence des gènes AS-impactés qui sont connectés entre eux, et les arbres de Steiner permettent d'identifier les acteurs dans le voisinage de ces gènes AS-impactés. Une analyse biologique des résultats de notre analyse des reconfigurations à l'aide des arbres de Steiner est discutée dans le chapitre suivant.



6

Discussion biologique

Nous avons présenté dans les chapitres précédents les méthodes développées au cours de la thèse ainsi que les résultats fournis par chacun des traitements que nous avons proposés. Les outils développés sont des outils de fouille de données qui permettent de mettre en évidence des fonctions biologiques ou des listes de gènes pour qu'une investigation biologique plus poussée puisse être réalisée. Nous discutons ainsi, dans ce chapitre, de l'interprétation biologique qui peut être faite sur les résultats des différentes analyses présentées dans cette thèse.

Nous allons tout d'abord faire un bref rappel sur les données présentées dans le chapitre 3.

Dans le chapitre 4, nous avons défini les termes révélés par les anti-sens et donné des informations sur le nombre de termes ainsi trouvés. Nous présentons maintenant les traitements bio-informatiques supplémentaires réalisés à partir de ces termes révélés par les anti-sens, pour interpréter et discuter leur signification biologique.

Ensuite, nous revenons sur les motifs de changement produits par notre analyse différentielle de réseaux présentée dans le chapitre 5. Pour interpréter les informations apportées par ces motifs de changement , nous proposons tout d'abord d'étudier si les interactions impactées par les anti-sens correspondent à des interactions biologiques connues. Nous proposons également d'étudier les processus biologiques concernés par ces motifs de changement en réalisant des tests d'enrichissement fonctionnel pour chaque motif. Enfin nous donnons les résultats de l'étude des reconfigurations des interactions entre gènes AS-impactés réalisée à l'aide des arbres de Steiner.

6.1 Contexte biologique	119
-----------------------------------	-----

6.2 Interprétation de l'analyse fonctionnelle différentielle	119
--	-----

6.2.1	Interprétation biologique	121
6.2.2	Évolution de l'expression des couples associés à un terme révélé par les anti-sens	126
6.2.3	Identification de sites de fixation connus	127
6.3	Interprétation de l'analyse différentielle de réseaux	131
6.3.1	Interprétation des résultats de l'analyse différentielle de réseau à l'aide de connaissances biologiques	132
6.3.2	Étude de l'enrichissement des motifs de changement	136
6.3.3	Sous-graphes AS-impactés	139

6.1 Contexte biologique

Les données utilisées lors de cette thèse sont des données de fruits de pommiers dont l'expression des gènes est mesurée au moment de la récolte (H) puis deux mois plus tard (60DAH). Pendant les deux mois, les fruits sont conservés dans des chambres froides.

Les facteurs post-récolte les plus importants qui influencent la perte de fermeté du fruit sont entre autres la température, l'atmosphère, l'humidité, le traitement au calcium et l'éthylène [Johnston *et al.*, 2002]. Essayer de prévenir le processus de maturation et préserver la qualité du fruit est l'intérêt principal du stockage dans une atmosphère contrôlée (bas en oxygène et haut en oxyde de carbone) et/ou dans des chambres froides. Dans les chambres froides, les pommes sont stockées dans des chambres de 0 à 3°C. Les basses températures influencent la biologie des pommes pendant la phase d'après récolte. Le stress induit par le froid provoque un désordre physiologique [Yang *et al.*, 2012]. Les basses températures dérèglent l'équilibre des dérivés réactifs de l'oxygène, entraînant son accumulation et un stress oxydatif [Chaparzadeh et Yavari, 2013]. Ainsi, dans notre cas où les fruits sont stockés pendant deux mois dans une chambre froide, les mécanismes principalement en jeu sont la maturation et les modifications de paroi cellulaire, les voies de signalisation du froid et la réponse à un stress froid, ainsi que la réponse à un stress oxydatif.

Rappel sur les données

Dans ce contexte, nous considérons l'expérience H comme étant le « contrôle » et l'expérience 60DAH le « traitement » de la maturation du fruit. Nous rappelons que les échantillons des expériences sont constitués de génotypes de pommes différents, mais que les échantillons sont appariés : à chaque échantillon de H correspond un échantillon de 60DAH.

L'étude a été menée à partir de l'ensemble des données du génome du pommier. Cependant, les analyses se concentrent sur les transcrits d'intérêt : les transcrits qui sont différentiellement exprimés entre H et 60DAH. Les transcrits sont considérés comme différentiellement exprimés si leur évolution est non-nulle, c'est-à-dire que les expressions moyennes des deux expériences diffèrent de plus d'un log. Par la suite, les analyses différentielles sont effectuées à partir des transcrits d'intérêt. Nous travaillons avec 1 625 transcrits d'intérêt : 931 transcrits sens et 624 transcrits anti-sens.

6.2 Interprétation de l'analyse fonctionnelle différentielle

L'analyse fonctionnelle différentielle permet d'identifier les termes révélés par les anti-sens. Les termes révélés par les anti-sens sont les termes de la Gene Ontology issus d'un enrichissement fonctionnel lorsqu'on utilise les données sens et anti-sens, mais qui n'apparaissent pas lorsqu'on utilise uniquement les données sens. La figure 6.1 rappelle la méthodologie appliquée afin d'identifier les termes révélés par les anti-sens ; nous indiquons de plus sur cette figure quels traitements bio-informatiques supplémentaires

nous avons réalisés afin d'interpréter l'information apportée par ces termes.

Nous effectuons trois traitements différents détaillés ci-après (figure 6.1). Le premier traitement est une interprétation biologique de ces termes par une analyse des fonctions associées aux termes. Le second traitement est l'observation du profil d'expression des transcrits d'intérêt associés aux termes révélés par les anti-sens. Le troisième traitement est l'analyse des régions supposées promotrices de ces transcrits afin de rechercher des sites de fixation de facteurs de transcription connus.

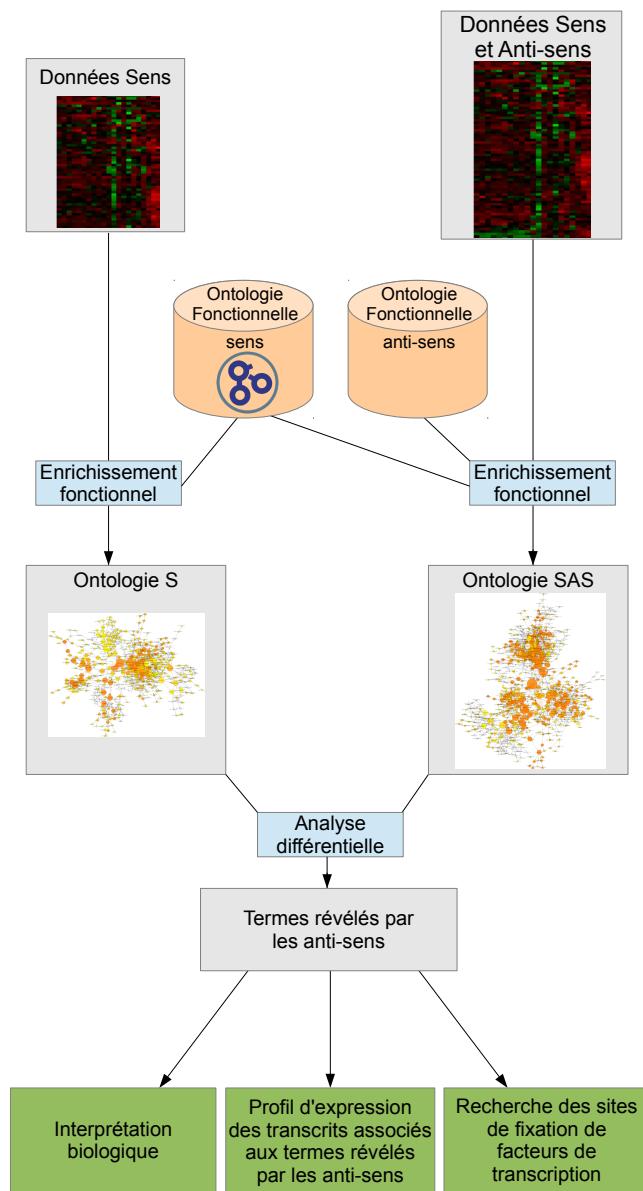


FIGURE 6.1 – Méthodologie de l'analyse fonctionnelle différentielle et traitements appliqués afin d'interpréter les résultats.

6.2.1 Interprétation biologique

Avant de regarder quels sont les termes révélés par les anti-sens, nous donnons dans la table 6.1 les résultats de tests d'enrichissement fonctionnel réalisés à partir des données sens uniquement. Il s'agit donc de l'analyse fonctionnelle classique cherchant à caractériser les processus caractérisant les gènes d'intérêt sens. Parmi cette liste on retrouve des termes liés à des processus impliqués dans la maturation du fruit comme les termes liés aux hormones (“flavonoid synthetic process”, “regulation of hormone levels”) et au transport (“nitrogen compound transport”, “ion transport”). Il est tout à fait logique de retrouver ce genre de termes fonctionnels dans le contexte de la maturation du fruit. L'enrichissement fonctionnel des données sens nous donne donc un résultat cohérent avec le contexte physiologique étudié.

TABLE 6.1 – Liste complète des termes enrichis par l'analyse des données sens. Pour chaque terme, nous indiquons la p-valeur associée (ce critère est utilisé pour trier la table) et le nombre de transcrits annotés par ce terme dans notre ensemble de transcrits d'intérêt sens.

GO catégorie	p-valeur	Nombre de transcrits
nitrogen compound transport	5.3124e-04	64
photosynthesis, light reaction	5.3124e-04	30
small molecule biosynthetic process	6.0977e-04	110
oxidation-reduction process	1.0328e-03	105
ion transport	1.0328e-03	91
photosynthesis	1.5814e-03	34
photosystem II assembly	2.5110e-03	19
anion transport	2.6387e-03	52
reductive pentose-phosphate cycle	5.7043e-03	7
generation of precursor metabolites and energy	5.7043e-03	52
photosynthesis, dark reaction	5.7043e-03	7
flavonoid biosynthetic process	5.7043e-03	19
response to karrikin	5.7043e-03	18
response to gibberellin	5.7043e-03	19
organonitrogen compound catabolic process	5.7043e-03	34
response to red light	6.2756e-03	14
organic acid transport	7.0877e-03	31
electron transport chain	7.4832e-03	14
carbon fixation	7.4832e-03	7
cysteine biosynthetic process	7.9422e-03	22
flavonoid metabolic process	7.9422e-03	19
regulation of hormone levels	8.5316e-03	38
carboxylic acid transport	8.5316e-03	30
cysteine metabolic process	8.6985e-03	22
response to chemical	9.1216e-03	207
response to UV	9.3830e-03	28
serine family amino acid biosynthetic process	9.4146e-03	22
steroid biosynthetic process	9.4146e-03	24
sulfur amino acid biosynthetic process	9.4146e-03	28
amino acid transport	1.0002e-02	27
nitrate transport	1.0002e-02	20
gibberellic acid mediated signaling pathway	1.0002e-02	13
response to oxygen-containing compound	1.0002e-02	138
alpha-amino acid metabolic process	1.0106e-02	46
steroid metabolic process	1.0106e-02	25
amide transport	1.0106e-02	16

TABLE 6.1 – Liste complète des termes enrichis par l’analyse des données sens. (suite)

GO catégorie	p-valeur	Nombre de transcrits
gibberellin mediated signaling pathway	1.0106e-02	13
hormone metabolic process	1.0106e-02	27
cellular response to gibberellin stimulus	1.1590e-02	13
sulfur compound biosynthetic process	1.1858e-02	40
response to endogenous stimulus	1.2674e-02	117
response to sucrose	1.3525e-02	22
response to disaccharide	1.4065e-02	22
alpha-amino acid biosynthetic process	1.4282e-02	35
organic anion transport	1.5335e-02	30
response to far red light	1.5335e-02	13
sulfur compound metabolic process	1.5726e-02	47
alcohol biosynthetic process	1.5763e-02	29
response to light stimulus	1.6344e-02	80
photosynthetic electron transport in photosystem I	1.6685e-02	8
inorganic anion transport	1.7261e-02	23
alcohol metabolic process	1.7261e-02	34
secondary metabolic process	1.7518e-02	48
amine catabolic process	1.7518e-02	15
cellular biogenic amine catabolic process	1.7518e-02	15
organic acid biosynthetic process	1.7518e-02	80
carboxylic acid biosynthetic process	1.7518e-02	80
response to organic substance	1.7629e-02	156
cellular glucan metabolic process	1.7629e-02	38
glucan metabolic process	1.7629e-02	38
serine family amino acid metabolic process	2.0056e-02	25
cellular homeostasis	2.2063e-02	24
sulfur amino acid metabolic process	2.2075e-02	29
response to radiation	2.2558e-02	82
organonitrogen compound metabolic process	2.2558e-02	109
response to fungus	2.4821e-02	46
response to stimulus	2.4821e-02	341
photosynthesis, light harvesting in photosystem I	2.4821e-02	3
cellular response to xenobiotic stimulus	2.4821e-02	3
xenobiotic metabolic process	2.4821e-02	3
response to acid chemical	2.4821e-02	96
cellular ion homeostasis	2.4821e-02	18
response to hormone	2.4821e-02	106
single-organism biosynthetic process	2.4821e-02	184
oligopeptide transport	2.4862e-02	14
peptide transport	2.4862e-02	14
photosynthetic electron transport chain	2.6829e-02	9
response to abiotic stimulus	2.8743e-02	157
cellular amino acid biosynthetic process	3.0481e-02	38
small molecule metabolic process	3.0481e-02	165
phenylpropanoid metabolic process	3.2419e-02	24
cellular cation homeostasis	3.2616e-02	17
regulation of ion homeostasis	3.2616e-02	2
positive regulation of trichoblast fate specification	3.2616e-02	2
positive regulation of plant epidermal cell differentiation	3.2616e-02	2
positive regulation of cell fate specification	3.2616e-02	2
response to nitrate	3.2616e-02	17
cellular amino acid metabolic process	3.5215e-02	53

TABLE 6.1 – Liste complète des termes enrichis par l'analyse des données sens. (suite)

GO catégorie	p-valeur	Nombre de transcrits
gibberellin catabolic process	4.2320e-02	3
diterpenoid catabolic process	4.2320e-02	3
amino acid import	4.2320e-02	12
single-organism transport	4.2346e-02	156
gibberellin biosynthetic process	4.4578e-02	8
polyamine catabolic process	4.8414e-02	9
gibberellin metabolic process	4.9155e-02	8

Dans le chapitre 4, nous avons proposé une analyse différentielle afin de savoir si, en effectuant un enrichissement fonctionnel avec l'ensemble des données sens et anti-sens, on obtient des termes nouveaux permettant de caractériser des processus mis en jeu dans notre contexte biologique mais qui sont sur-représentés seulement quand on prend en compte les données anti-sens. Pour répondre à cette question, nous avons calculé les termes révélés par les anti-sens.

La table 6.2 liste les 104 termes révélés par les anti-sens identifiés dans les données pommier. Parmi cette liste, on retrouve également des termes liés à des processus impliqués dans la maturation du fruit. On retrouve ainsi des termes liés à la paroi cellulaire (“cell wall organization or biogenesis”, “multidimensional cell growth”) et à la réponse osmotique (“water transport”, “response to osmotic stress”). Lors de la maturation du fruit, les fruits continuent de vivre et des échanges se produisent au niveau de la paroi, on observe également une modification de la pression osmotique [Prasanna *et al.*, 2007].

Le plus surprenant est de trouver également des termes liés à la réponse contre le froid (“response to temperature stimulus”, “response to cold”). Les fruits étant stockés dans des chambres froides lors de leur maturation, il apparaît en effet étonnant de ne pas trouver ces termes dans une analyse comportant uniquement les données sens.

En comparant les deux enrichissements fonctionnels sur les données sens uniquement d'un côté et sur les données sens et anti-sens de l'autre, on découvre ainsi des termes fonctionnels pertinents pour le contexte biologique étudié. De plus, cela suggère que les transcrits anti-sens jouent un rôle particulier dans les processus biologiques ainsi mis en évidence : fonctions liées à la paroi cellulaire, à la pression osmotique et à la réponse à un stress de froid. Ces fonctions biologiques sont cohérentes avec le contexte de la maturation du fruit stocké en chambres froides et il est intéressant, voire surprenant, d'obtenir ces informations uniquement lors de l'ajout des données anti-sens. Dans la table 6.2, nous indiquons combien de transcrits, sens et anti-sens, étiquettent chaque terme révélé par les anti-sens. Notre analyse fournit bien sûr les noms de ces transcrits et nous allons proposer, dans la section suivante, une exploration plus approfondie de l'expression de ces transcrits.

TABLE 6.2 – Liste complète des termes révélés par les anti-sens. Pour chaque terme, nous indiquons la p-valeur associée (ce critère est utilisé pour trier la table) et le nombre de transcrits annotés par ce terme dans notre ensemble de transcrits d'intérêt.

GO catégorie	p-valeur	Nombre de transcrits
single-organism localization	2.5178e-04	285
response to blue light	5.5327e-04	25

TABLE 6.2 – Liste complète des termes révélés par les anti-sens. (suite)

GO catégorie	p-valeur	Nombre de transcrits
root hair cell development	5.5327e-04	36
root hair elongation	5.6530e-04	35
regulation of trichoblast fate specification	6.2544e-04	4
regulation of plant epidermal cell differentiation	6.2544e-04	4
regulation of cell fate specification	6.2544e-04	4
water transport	1.0017e-03	32
fluid transport	1.0017e-03	32
positive regulation of cell fate commitment	1.0769e-03	4
brassinosteroid biosynthetic process	1.6036e-03	22
phytosteroid biosynthetic process	1.6036e-03	22
response to salt stress	2.0084e-03	100
response to lipid	2.1347e-03	94
transport	2.2339e-03	303
phytosteroid metabolic process	2.2339e-03	22
brassinosteroid metabolic process	2.2339e-03	22
regulation of proton transport	2.9591e-03	16
regulation of cell fate commitment	3.2819e-03	4
external encapsulating structure organization	3.2819e-03	87
response to temperature stimulus	3.3930e-03	108
response to osmotic stress	3.3930e-03	103
cell wall organization or biogenesis	4.2486e-03	107
establishment of localization	4.3669e-03	306
multidimensional cell growth	4.6029e-03	23
organic hydroxy compound biosynthetic process	4.8442e-03	81
cation transport	4.9288e-03	99
cell wall organization	4.9288e-03	83
sterol metabolic process	5.1595e-03	32
sterol biosynthetic process	6.0117e-03	31
positive regulation of cell differentiation	7.5887e-03	4
nicotinamide nucleotide metabolic process	7.9948e-03	57
pyridine nucleotide metabolic process	8.1151e-03	57
cellular response to nitrate	9.6834e-03	4
pyridine-containing compound metabolic process	1.0042e-02	57
developmental cell growth	1.0108e-02	46
response to carbohydrate	1.1066e-02	51
pigment accumulation in response to UV light	1.1129e-02	21
pigment accumulation in tissues in response to UV light	1.1129e-02	21
pigment accumulation in tissues	1.1129e-02	21
anthocyanin accumulation in tissues in response to UV light	1.1129e-02	21
glycine decarboxylation via glycine cleavage system	1.1129e-02	5
pigment accumulation	1.1129e-02	21
pigmentation	1.1129e-02	21
hormone biosynthetic process	1.1601e-02	34
localization	1.1690e-02	314
regulation of biological quality	1.1758e-02	126
cellular protein complex assembly	1.1893e-02	59
trichoblast maturation	1.3064e-02	44
root hair cell differentiation	1.3064e-02	44
cell maturation	1.3214e-02	44
polysaccharide metabolic process	1.4240e-02	90
response to alcohol	1.5038e-02	79
Golgi organization	1.6037e-02	34

TABLE 6.2 – Liste complète des termes révélés par les anti-sens. (suite)

GO catégorie	p-valeur	Nombre de transcrits
peptide cross-linking	1.6037e-02	2
peptide cross-linking via L-cystine	1.6037e-02	2
nicotinate transport	1.6037e-02	2
N-methylnicotinate transport	1.6037e-02	2
quaternary ammonium group transport	1.6037e-02	2
second-messenger-mediated signaling	1.6391e-02	12
regulation of root morphogenesis	1.8440e-02	4
trichoblast differentiation	1.9012e-02	44
secondary metabolite biosynthetic process	2.0762e-02	46
peptidyl-cysteine modification	2.1434e-02	4
cytokinin catabolic process	2.1434e-02	4
hormone catabolic process	2.1434e-02	4
glycosyl compound metabolic process	2.2747e-02	66
cellular modified amino acid metabolic process	2.3978e-02	20
transmembrane transport	2.4361e-02	88
response to desiccation	2.7329e-02	9
plant-type cell wall cellulose metabolic process	2.8846e-02	9
developmental maturation	2.8846e-02	61
energy quenching	2.9392e-02	4
nonphotochemical quenching	2.9392e-02	4
organic hydroxy compound metabolic process	3.0436e-02	91
developmental growth involved in morphogenesis	3.2066e-02	54
long-day photoperiodism, flowering	3.2558e-02	7
response to light intensity	3.2558e-02	39
response to vitamin B1	3.2558e-02	2
auxin metabolic process	3.2558e-02	19
developmental growth	3.2899e-02	60
regulation of ion transport	3.5655e-02	27
cellular response to ozone	3.5962e-02	3
plant-type cell wall organization	3.5962e-02	37
response to red or far red light	3.5962e-02	48
response to cold	3.5962e-02	65
long-day photoperiodism	3.5962e-02	7
oxoacid metabolic process	3.5962e-02	209
cellular chemical homeostasis	3.6502e-02	27
organic acid metabolic process	3.6502e-02	209
root epidermal cell differentiation	3.7766e-02	44
endomembrane system organization	3.8916e-02	35
cellular metabolic compound salvage	4.0913e-02	26
anion transmembrane transport	4.1764e-02	23
hyperosmotic response	4.2619e-02	38
metal ion transport	4.2619e-02	63
response to high light intensity	4.2880e-02	31
root morphogenesis	4.3141e-02	52
plant epidermal cell differentiation	4.3141e-02	44
oxidoreduction coenzyme metabolic process	4.3698e-02	59
single-organism metabolic process	4.5237e-02	505
starch metabolic process	4.8270e-02	29
organonitrogen compound biosynthetic process	4.9633e-02	107
protein complex biogenesis	4.9633e-02	65

6.2.2 Évolution de l'expression des couples associés à un terme révélé par les anti-sens

Dans la table 6.2, nous listons les termes révélés par les anti-sens et le nombre de transcrits associés à chacun des termes fonctionnels. Ce nombre de transcrits est le nombre de transcrits d'intérêt, sens ou anti-sens, qui sont différentiellement exprimés entre H et 60DAH. Nous proposons ici une étude du profil d'expression de chacun de ces transcrits d'intérêt. Nous avons expliqué dans le chapitre 1 que le transcrit anti-sens est complémentaire de son transcrit sens avec lequel il peut s'hybrider. Dans nos données nous avons observé peu de gènes pour lesquels à la fois le transcrit sens et le transcrit anti-sens sont différentiellement exprimés. Néanmoins, dans l'analyse qui suit, il nous a paru intéressant de considérer à la fois un transcrit différentiellement exprimé et son complémentaire.

Rappelons que nous avons identifié 1 625 transcrits d'intérêt (931 sens et 694 anti-sens) avec 200 gènes pour lesquels à la fois le sens et l'anti-sens sont différentiellement exprimés. En considérant les couples (sens, anti-sens) dont au moins un élément est différentiellement exprimé, on obtient donc 1 425 couples soit 2 850 transcrits. La table 6.3 dénombre les couples de transcrits sens et anti-sens de gènes d'intérêt selon l'évolution des transcrits. On ajoute ainsi aux transcrits d'intérêt les transcrits complémentaires qui ont une évolution nulle. De la table 6.3 on retire plusieurs remarques. On remarque ainsi que la majorité des sondes qui évoluent ont une évolution négative entre H et 60DAH. On remarque également que seulement quatre couples voient leur transcrit sens avoir une évolution opposée au transcrit anti-sens. Enfin on note que beaucoup de couples ont la même évolution mais majoritairement l'un des transcrits n'évolue pas.

TABLE 6.3 – Évolutions des couples de sondes des transcrits d'intérêt. Cette table dénombre les sondes sens (S) et anti-sens (AS) en fonction de leur évolution (-1, 0, 1). Seuls les transcrits d'intérêt sont ici présents, c'est pourquoi il n'y a aucun transcrit où le sens et l'anti-sens ont une évolution nulle.

		S		
		-1	0	1
AS		-1	136	371
		0	420	0
		1	3	123
				60

Nous nous intéressons donc ici aux valeurs d'expression des transcrits associés à un terme révélé par les anti-sens. Puisqu'il s'agit de termes révélés par les anti-sens, nous trouvons nécessairement parmi ces transcrits un nombre relativement important d'anti-sens qui ont une évolution non-nulle. Nous avons donc choisi de représenter dans un diagramme l'évolution à la fois des sondes sens et anti-sens associées à un terme révélé par les anti-sens. Cette visualisation nous permet de rechercher des comportements significatifs en explorant l'évolution de ces sondes. La visualisation trie les évolutions en fonction de l'évolution du transcrit anti-sens dans un premier temps puis du transcrit sens.

Ainsi, la figure 6.2 montre l'évolution des gènes étiquetés par deux termes révélés par les anti-sens : le terme "response to cold" et le terme "hyperosmotic response". Nous sélectionnons ces deux termes

spécifiquement puisqu'ils sont des termes révélés par les anti-sens qui sont des spécifications de deux autres termes, respectivement "response to temperature stimulus" et "response to osmotic stress" qui ont une p-valeur faible (de l'ordre de 10^{-3}) mais qui sont rattachés à trop de gènes pour que les images soient lisibles dans ce document. Sur ces images, on peut ainsi comprendre pourquoi ces termes sont révélés par les anti-sens : tous les anti-sens qui évoluent alors que leur sens complémentaire n'évolue pas, sont des acteurs supplémentaires pris en compte dans l'analyse fonctionnelle sens et anti-sens. C'est leur présence qui a conduit à obtenir ce terme comme sur-représenté lors du test d'enrichissement.

Globalement, il existe plus de sondes qui ont une évolution négative entre H et 60DAH que de sondes qui évoluent positivement (table 6.3). On observe le même phénomène lorsqu'on regarde un ensemble de gènes annotés par un terme révélé par les anti-sens. De même, il existe peu de couples pour lesquels l'évolution du sens est opposée à l'évolution de l'anti-sens.

L'image produite indique, en plus de l'évolution des couples, l'orthologue correspondant. L'outil développé détecte les orthologues identiques et les signale en surlignant les orthologues identiques d'une même couleur. Placer l'évolution des couples en parallèle de l'orthologue permet de montrer les limites de la prédiction de la relation d'orthologie. On remarque ainsi que plusieurs gènes du pommier possèdent le même orthologue avec *Arabidopsis*. Si la prédiction est bonne, on s'attend donc à ce que ces gènes réagissent de la même manière dans les différentes conditions. On observe effectivement que la plupart des gènes qui possèdent le même orthologue ont la même évolution. Sachant que l'orthologie est calculée à partir du transcript sens, on observe également que la plupart des gènes qui ont un même orthologue ont leur transcript sens qui évoluent de la même manière.

Cependant, on observe également que certains gènes n'évoluent pas de la même manière et certains ont même une évolution opposée. Sur la figure 6.2 on observe en effet que les couples de gènes annotés "hyperosmotic response" qui ont un même orthologue ont la même évolution, à l'exception des couples dont l'orthologue est AT4G33000 (fond violet) où le transcript sens d'un couple évolue négativement alors que le transcript sens de l'autre couple évolue positivement ; le transcript anti-sens du premier couple n'évolue pas alors que le transcript anti-sens du second évolue positivement. L'annotation structurale des gènes et/ou la relation d'orthologie prédite peuvent être remises en cause dans ce cas.

Nous ne donnons ici que deux figures mais l'outil que nous avons développé permet de générer automatiquement ces figures pour chacun des termes révélés par les anti-sens. Cette visualisation permet de lister les transcrits associés à un terme ontologique en les regroupant par profil d'évolution et donc ainsi de voir s'il existe une évolution dominante pour ce terme.

6.2.3 Identification de sites de fixation connus

Afin d'approfondir l'analyse fonctionnelle différentielle que nous avons proposée, nous avons mené une étude des séquences promotrices des gènes étiquetés par un terme révélé par les anti-sens. Les séquences promotrices des gènes sont les séquences qui précèdent la partie transcrive du gène et qui permet aux facteurs de transcription de se fixer sur l'ADN. On connaît les séquences qui permettent de fixer certains

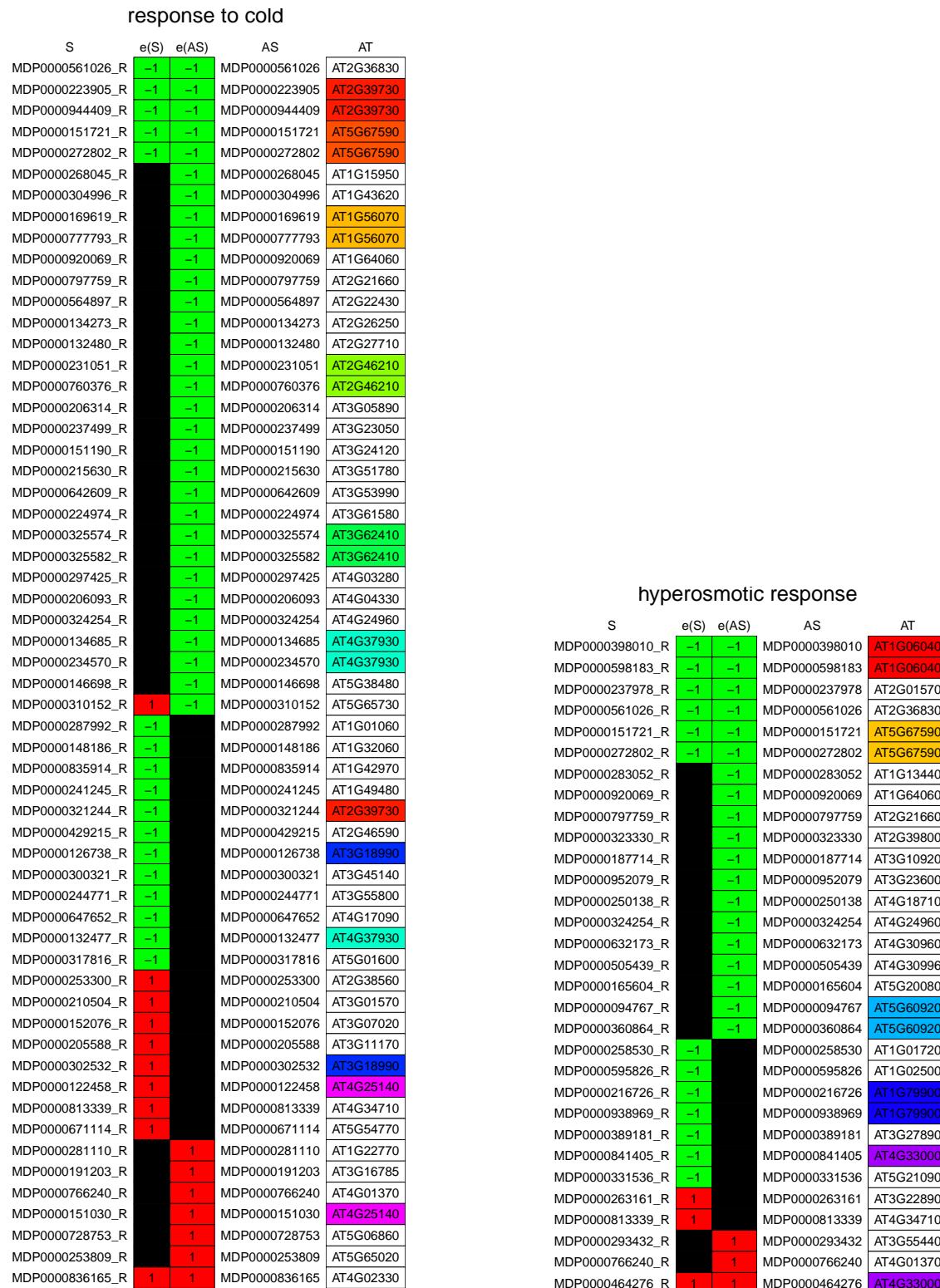


FIGURE 6.2 – Évolutions des gènes étiquetés par des termes révélés par les anti-sens. Les sondes sens sont situées à gauche, les sondes anti-sens à droite. L'évolution des sondes est indiquée au milieu ($e(S)$) pour les sens, $e(AS)$ pour les anti-sens), soit elle est négative (verte), nulle (noire) ou positive (rouge) entre H et 60DAH. À l'extrême droite l'orthologue d'Arabidopsis est indiqué, lorsque deux orthologues sont identiques ils ont une couleur de fond autre que blanche identique.

facteurs de transcription et, pour certains facteurs de transcription, on connaît également dans quelles voies biologiques ils sont impliqués. En analysant les sites de fixation de facteurs de transcription connus, on peut ainsi prédire les voies biologiques dans lesquelles le transcript peut intervenir.

L'analyse des promoteurs se fait en recherchant des motifs nucléiques particuliers dans la région promotrice du gène. La région promotrice du gène se situe en amont du gène, c'est entre autre ici que se fixent les facteurs de transcription afin d'initier la transcription du gène. Ici, un motif est une séquence nucléique, pouvant, par exemple, prendre la forme (A | T) GCC qui signifie que la séquence débute par de l'adénine ou de la thymine, suivi par de la guanine et deux cytosines. Cette analyse peut s'effectuer grâce à l'outil en ligne RSAT [Medina-Rivera *et al.*, 2015]. Cet outil permet, parmi d'autres fonctionnalités, de rechercher des motifs sur-représentés dans un ensemble de séquences, ou de rechercher un motif particulier dans cet ensemble.

La recherche de motifs sur-représentés s'effectue par l'assemblage d'oligomères de tailles variables (de 1 à 8 nucléotides). Un modèle mathématique basé sur le modèle de Markov, associé à un organisme végétal permet de dire si un motif est sur-représenté. La recherche peut s'effectuer sur un brin spécifique (la séquence donnée) ou sur les deux brins (la séquence donnée et son complémentaire inversé). Le résultat de la recherche est un ensemble de motifs sur-représentés. On recherche ensuite l'emplacement de ces motifs dans les séquences et on peut ainsi comparer ces emplacements afin de savoir si les sites de fixation d'un même facteur de transcription sont situés aux mêmes endroits de la région promotrice des différents gènes.

Nous nous intéressons plus particulièrement à la fonction "response to cold" en recherchant, dans les séquences promotrices de ces gènes, des motifs de site de fixation de facteurs de transcription connus pour intervenir lors d'un stress froid, comme ABRE ou DRE [Yamaguchi-Shinozaki et Shinozaki, 2005]. Nous travaillons sur des séquences de pommier, pour lesquels les orthologues d'Arabidopsis ont été annotés par la fonction "response to cold", nous vérifions donc si ces motifs apparaissent sur les sens. Parmi les 56 gènes d'intérêt qui sont étiquetés "response to cold", on a retrouvé ABRE dans 26 séquences promotrices, ce qui nous confirme l'implication de ces gènes dans un mécanisme de réponse à un stress de froid.

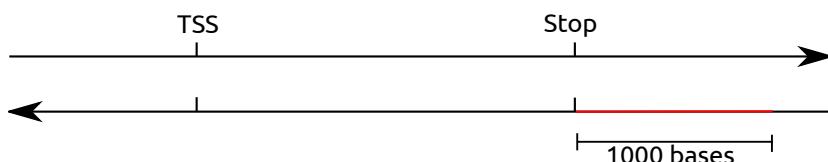


FIGURE 6.3 – Emplacement putatif de la région promotrice de l'anti-sens (rouge). Le gène est situé sur le brin supérieur, la partie transcrrite du gène est comprise entre le site d'initiation de la transcription (TSS) et le codon stop (Stop).

Nous appliquons maintenant la recherche de sites de fixation aux transcrits anti-sens, afin d'identifier s'ils ont des sites de fixation communs et si ces sites de fixation sont connus. En effet, la séquence transcrrite de l'anti-sens est située sur le brin opposé de la séquence du sens, ce qui implique que le sens

de lecture du brin est opposé également et que le sens et l'anti-sens ne partagent donc pas leur séquence promotrice. Le début et la fin de la séquence transcrrite est connue pour le sens, mais nous ne savons pas exactement où débute la transcription de l'anti-sens. La figure 6.3 illustre où se situe la zone que nous considérons comme promotrice pour l'anti-sens. Ne sachant pas exactement où se trouve le transcription start site de l'anti-sens, nous considérons qu'il doit se situer en amont de l'endroit où se termine la région transcrrite du sens. Afin de s'assurer que la région promotrice soit dans notre analyse, nous avons extrait la séquence composée des 1 000 bases précédant l'emplacement du codon stop sur le brin complémentaire (en rouge dans la figure 6.3).

Nous avons restreint l'analyse aux anti-sens qui évoluent négativement entre les deux expériences H et 60DAH, c'est-à-dire les anti-sens pour lequel le niveau d'expression est plus élevé à la récolte que deux mois plus tard. Nous avons ainsi travaillé sur un ensemble de 30 séquences.

Après avoir recherché les motifs sur-représentés dans l'ensemble des séquences, nous identifions les sites de fixations en interrogeant la base de données AGRIS [Davuluri *et al.*, 2003] qui regroupe un ensemble de motifs de sites de fixation pour *Arabidopsis*. Pour chaque motif trouvé dans AGRIS, nous recherchons ensuite dans quelles séquences et à quelles positions ces motifs se trouvent. Les motifs d'AGRIS sont généralement plus courts que ceux trouvés à partir des séquences, c'est pourquoi nous recherchons ensuite les motifs de la base et les motifs d'origine. C'est-à-dire que lorsqu'un motif est retrouvé dans la base, nous recherchons dans les séquences à la fois le motif trouvé, et le motif associé au site de fixation identifié dans la base. Par exemple, le motif TACGTGGCGC est sur-représenté dans les séquences, et le motif pour ABRE est (C | T) ACGTGGC et se retrouve donc dans le motif sur-représenté, on recherchera donc où se situent dans les séquences le motif d'ABRE ((C | T) ACGTGGC) et le motif sur-représenté (TACGTGGCGC).

La figure 6.4 montre le résultat de l'analyse pour les anti-sens “response to cold” évoluant négativement entre les deux expériences. Plusieurs motifs sont sur-représentés dans l'ensemble des séquences, mais peu de ces motifs sont listés dans AGRIS. Seuls “RAV1” et “Bellringer/replumless/pennywise BS1 IN AG” ont été identifiés. Le fait que le site d'initiation de la transcription de l'anti-sens n'est pas identifié nous empêche de voir s'il y a un alignement de certains motifs sur les séquences. Même si peu de sites de fixation connus ont été identifiés et qu'on ne voit pas s'il existe une position préférentielle des motifs sur les séquences, nous notons qu'il existe plusieurs motifs sur-représentés dans les régions putativement promotrices des anti-sens. Il est donc intéressant de regarder ces motifs de plus près : il peuvent être des sites de fixation pour des facteurs de transcription spécifiques aux anti-sens. Pour poursuivre cette étude, il faudra cependant connaître précisément l'emplacement du site d'initiation de la transcription de l'anti-sens. Il faudrait également avoir des données d'expression pour une plus grande diversité de conditions physiologiques pour s'assurer de la co-régulation de ces anti-sens et donc de leur dépendance aux mêmes facteurs de transcription.

Avec le nouveau séquençage du génome [Daccord *et al.*, 2017], il sera plus facile d'analyser les séquences promotrices des anti-sens. Ce nouveau séquençage permettra d'avoir une séquence ADN de meilleure qualité. De plus, l'IRHS qui a réalisé ce nouveau séquençage, a également réalisé des analyses

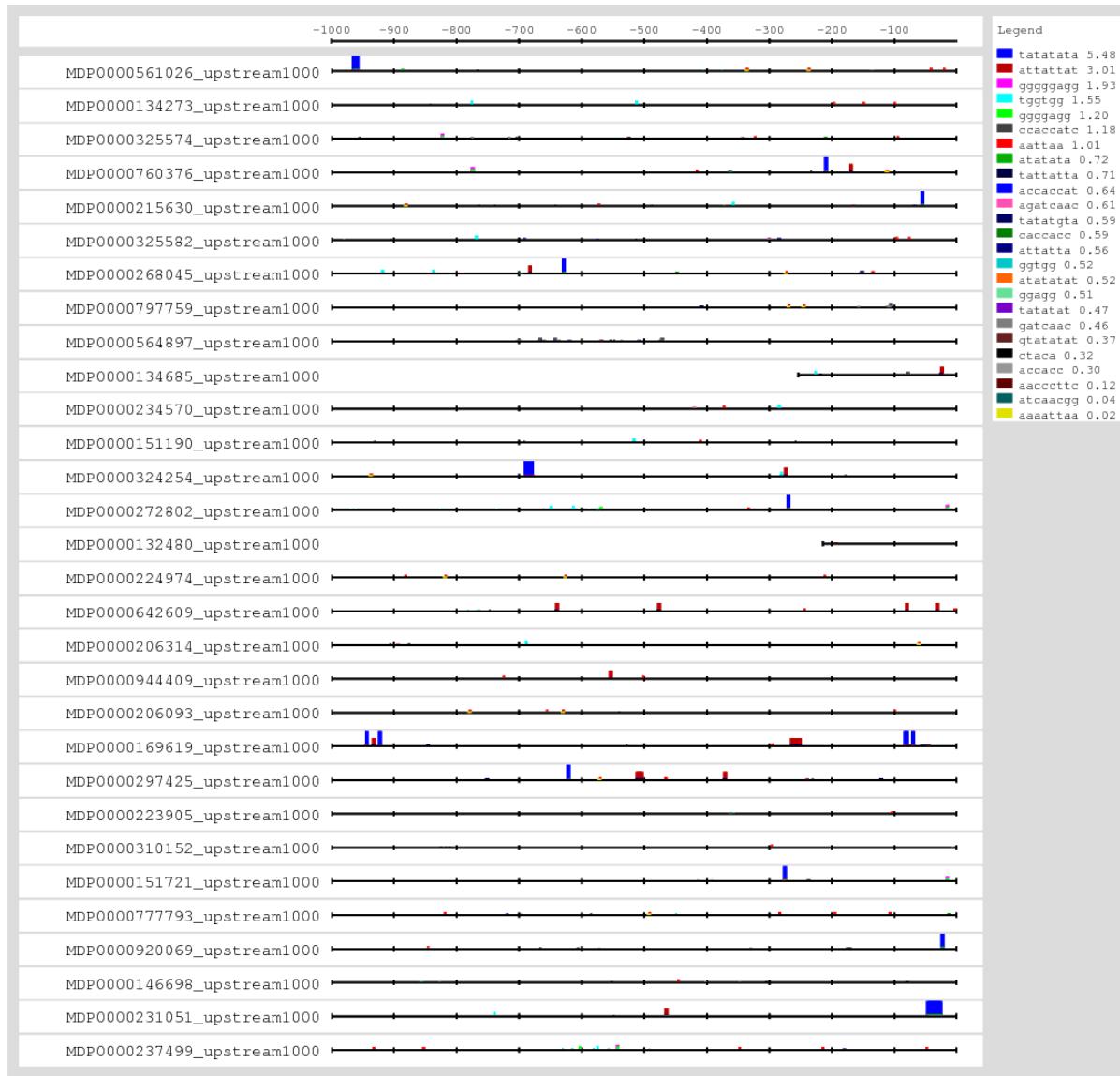


FIGURE 6.4 – Résultats de l'analyse des promoteurs des anti-sens associés à la fonction “response to cold” et évoluant négativement entre H et 60DAH.

RNA-Seq sur les transcrits anti-sens. Ce séquençage orienté des transcrits va permettre de connaître exactement la séquence correspondant au transcrit anti-sens. Cette séquence permet ainsi d'identifier où se situe le site d'initiation de la transcription de l'anti-sens, et donc d'identifier plus précisément la séquence promotrice de l'anti-sens.

6.3 Interprétation de l'analyse différentielle de réseaux

L'analyse différentielle de réseaux que nous avons présentée dans le chapitre 5 permet d'identifier les gènes AS-impactés et les motifs de changement. Les gènes AS-impactés sont des gènes pour lesquels l'intégration des données anti-sens impacte leur voisinage dans un cœur de réseau.

La figure 6.5 rappelle la méthodologie appliquée afin d’identifier les gènes AS-impactés et indique quels traitements bio-informatiques supplémentaires nous avons réalisés afin d’interpréter ces termes. D’abord nous interprétons les résultats de l’analyse différentielle grâce à la comparaison des interactions inférées avec des réseaux biologiques connus. Ensuite nous réalisons un test d’enrichissement fonctionnel des motifs de changement. Enfin nous étudions la reconfiguration, dans le réseau SAS, des interactions entre gènes AS-impactés, à l’aide des arbres de Steiner.

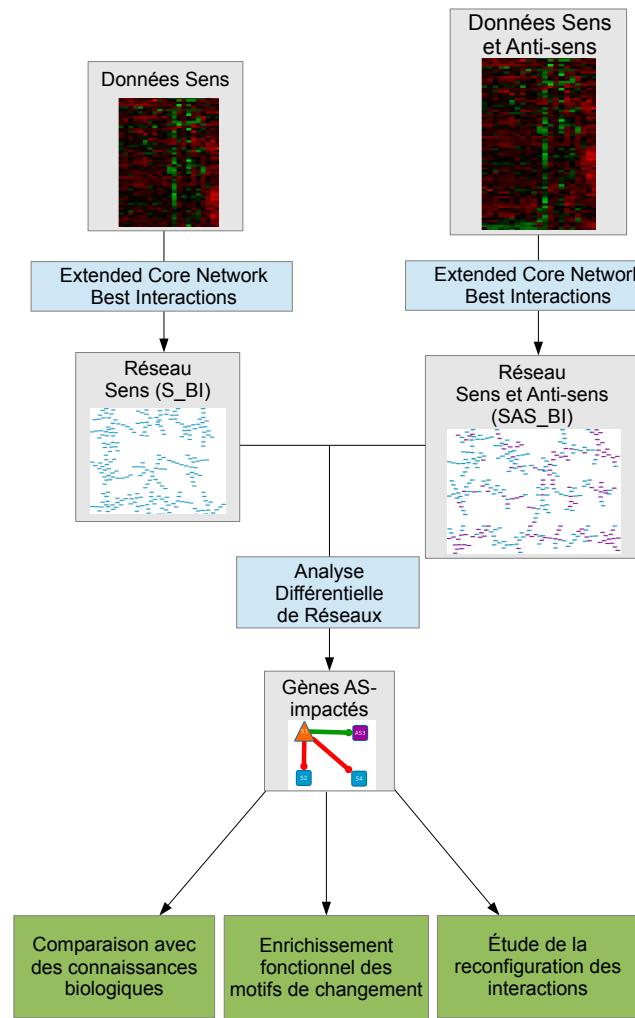


FIGURE 6.5 – Méthodologie de l’analyse différentielle de réseaux et traitements appliqués afin d’interpréter les résultats.

6.3.1 Interprétation des résultats de l’analyse différentielle de réseau à l’aide de connaissances biologiques

Un réseau de gènes est un graphe qui permet de représenter des interactions entre des gènes. Comme expliqué dans le chapitre 2, ces interactions observées au niveau transcriptomique peuvent traduire des

interactions effectives à différents niveaux dans la cellule : interaction facteur de transcription-gène, interaction protéine-protéine ou interaction métabolite-métabolite. Les interactions au niveau protéique sont des interactions physiques entre les protéines qui peuvent donc figurer dans les interactions du réseau de gènes.

Nous proposons maintenant de comparer les interactions inférées et les interactions participant à un motif de changement avec des interactions protéine-protéine connues et validées (figure 6.6).

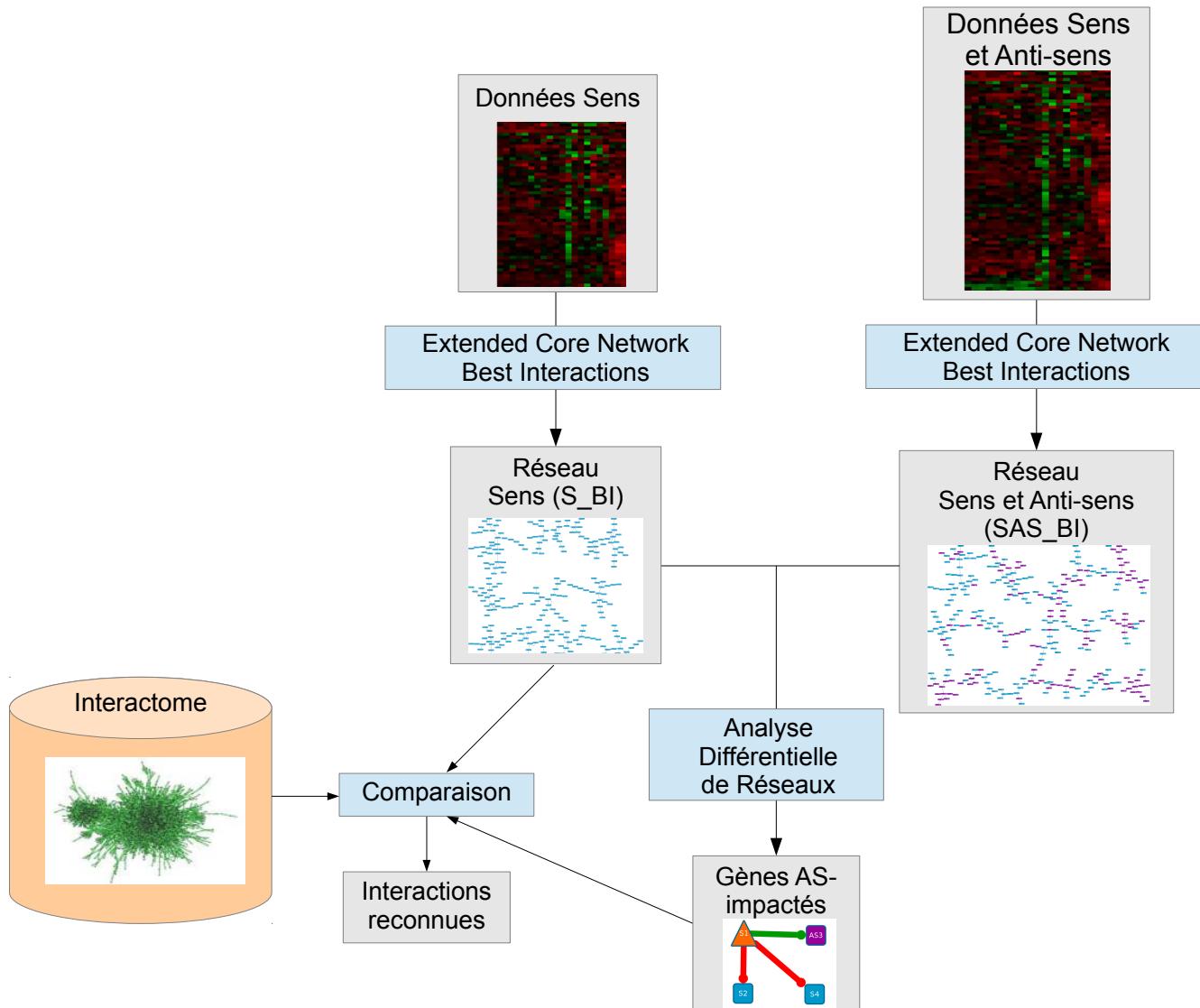


FIGURE 6.6 – Méthodologie de la comparaison des réseaux inférés avec l'interactome.

L'interactome est un graphe non-orienté qui représente une partie des interactions connues entre les protéines d'un organisme. Les nœuds du graphes représentent les protéines, et les arêtes sont les différentes interactions entre les protéines. Il existe plusieurs types d'interactions représentées dans un interactome : on peut ainsi représenter un lien physique où les deux protéines ont une réelle interaction biochimique mais on peut également représenter un lien bibliographique où les deux protéines sont ci-

tées dans un même document.

À l'heure actuelle il n'existe pas d'interactome disponible pour le pommier, nous utilisons donc l'interactome d'*Arabidopsis thaliana* [Consortium, 2011] et ses orthologues afin d'intégrer ces connaissances biologiques dans notre comparaison. Néanmoins, certains gènes d'intérêt utilisés dans notre étude ne se retrouvent pas dans l'interactome, et donc nous comparons l'interactome avec le sous-graphe du réseau inféré ne contenant que les gènes du pommier ayant un orthologue avec *Arabidopsis*.

Il existe plusieurs versions de l'interactome d'*Arabidopsis*. Nous utilisons la version AI-1. Les interactions de AI-1 proviennent des interactions protéine-protéine de l'ensemble principal (main) et des répétitions (repeat). Ces deux ensembles représentent les interactions biochimiques entre les protéines. On retrouve également des interactions provenant de la littérature dans AI-1. Cela forme ainsi un graphe de 4 866 noeuds et 11 374 interactions.

Nous souhaitons ainsi savoir quelles sont, parmi les interactions identifiées par notre méthode, celles qui correspondent à des relations connues et validées entre protéines. Puisque nous comparons notre réseau avec l'interactome, nous considérons ici uniquement le réseau S puisque seuls les transcrits sens peuvent produire une protéine. Nous voulons également identifier les interactions qui se retrouvent dans l'interactome et qui appartiennent à un motif de changement : ces interactions sont des interactions qui ne seront plus représentées dans le réseau SAS mais qui seront observées au niveau protéique.

Pour comparer les deux graphes, nous recherchons toutes les interactions du réseau inféré dans l'interactome. Puisque nous inférons un cœur de réseau avec notre méthode ECN et un taux d'acceptation de 0.05, nous autorisons que l'interaction dans l'interactome ne soit pas une interaction directe mais un chemin formé de plusieurs liens. Ainsi, lorsqu'une interaction $g_1 \rightarrow g_2$ est présente dans le réseau, on recherche s'il existe un chemin entre g_1 et g_2 dans l'interactome. Si on observe le chemin $g_1 - g_3 - g_2$ dans l'interactome, alors on considère que l'interaction $g_1 \rightarrow g_2$ du réseau se retrouve dans l'interactome. Cette décision peut être discutée et on pourrait restreindre la longueur du chemin recherché dans l'interactome. Dans notre cas, le chemin le plus long entre deux gènes g_1 et g_2 contient 4 gènes, g_1 et g_2 compris, ce qui reste donc un chemin de taille raisonnable. Nous n'avons retrouvé aucune interaction directe $g_1 - g_2$ dans l'interactome. Ceci n'est pas étonnant puisque le cœur de réseau ne représente que les interactions les plus fortes et cela ne correspond pas obligatoirement à une interaction physique entre les protéines traduites à partir des transcrits.

Nous avons donc réalisé cette comparaison uniquement avec le réseau S de l'expérience 60DAH. Parmi les 1 361 interactions du réseau S, nous en retrouvons seulement 47 dans l'interactome. Cette observation est normale puisque les interactions représentées dans un réseau de gènes ne sont pas censées toutes se retrouver au niveau des protéines, les interactions présentes dans l'interactome ne sont pas non plus censées se produire tout le temps dans tous les organes de la plante, et le cœur de réseau est une représentation très restreinte des interactions transcriptomiques. Cependant, nous pouvons noter que certaines de ces interactions sont impliquées dans un motif de changement. En effet, nous avons identifié 308 motifs de changements grâce à notre analyse différentielle, et parmi les 366 interactions du réseau S qui sont impliquées dans un motif de changement, on en retrouve 15 dans l'interactome. Nous avons donc



FIGURE 6.7 – Comparaison des interactions inférées avec les interactions de l’interactome. L’image représente le cœur de réseau inféré par Extended Core Network à partir des données sens dans lequel les interactions de l’interactome sont mises en valeur. Un lien vert en gras représente un lien que l’on retrouve dans l’interactome. Un lien rouge en gras représente un lien que l’on retrouve dans l’interactome et qui est dans un motif de changement.

47 interactions du réseau S qui se retrouvent dans l'interactome, dont 15 d'entre elles sont impliquées dans un motif de changement.

Lorsqu'on retrouve une interaction du cœur de réseau ECN dans l'interactome, cela signifie que l'interaction observée au niveau transcriptomique entre les deux acteurs sens se retrouve également au niveau protéique. Si cette interaction est également dans un motif de changement, cela signifie qu'il existe un transcrit anti-sens avec lequel un des deux acteurs sens partage une plus grande information mutuelle. Sous l'hypothèse d'une action des anti-sens dans le PTGS, cela peut signifier que l'anti-sens empêche l'interaction protéine-protéine en inhibant l'expression d'une des deux protéines. C'est pourquoi l'information est importante et nécessite une validation biologique. Cette comparaison permet donc de mettre en évidence des interactions intéressantes à étudier biologiquement.

La figure 6.7 montre le résultat graphique de la comparaison pour l'expérience 60DAH. On observe que peu d'interactions du réseau de gènes se retrouvent dans l'interactome. Cette visualisation permet néanmoins de situer où se trouvent ces interactions et de différencier les interactions qui participent à un motif de changement.

La nouvelle version du génome du pommier [Daccord *et al.*, 2017] et une meilleure caractérisation des orthologues avec *Arabidopsis* pourront fournir des résultats plus précis. Une meilleure association entre un gène du pommier et un gène d'*Arabidopsis* identifiera de manière plus exacte les interactions présentes dans l'interactome.

6.3.2 Étude de l'enrichissement des motifs de changement

Pour compléter l'analyse des résultats de l'analyse différentielle de réseaux de l'expérience 60DAH, nous proposons maintenant d'étudier plus en détail les fonctions biologiques associées aux motifs de changement.

L'analyse différentielle de réseaux a identifié 308 motifs de changement dans l'expérience 60DAH. Chaque motif de changement contient un gène AS-impacté, ses voisins directs dans le réseau S_BI, les anti-sens impactants et leurs voisins directs dans le réseau SAS_BI. Puisqu'un motif de changement est un ensemble de gènes produit par notre analyse, il est intéressant de savoir si des fonctions biologiques sont sur-représentées dans cet ensemble. Les motifs de changement sont formés de cinq gènes en moyenne. Vu que la taille de cet ensemble est relativement petit comparée à la taille du génome, nous réalisons l'enrichissement fonctionnel, sur chaque motif de changement, en utilisant la GO slim qui ne contient que 14 catégories. Ainsi, chaque test consiste à considérer l'ensemble des gènes d'un motif et à effectuer le test d'enrichissement que pour les catégories de la GO slim.

Parmi les 308 motifs de changement que nous avons identifiés, 72 motifs de changements sont enrichis par des catégories de la GO slim comme “developmental processes”, “response to abiotic or biotic stimulus”, “response to stress” et “transport” (voir Annexe A Liste des motifs de changements enrichis de 60DAH).

Les 72 motifs de changement enrichis forment un ensemble de 291 gènes du pommier. Ces gènes du

pommier sont associés à 209 homologues avec *Arabidopsis*. Il est intéressant de noter qu'en regardant les fonctions biologiques putatives des gènes dans les annotations du TAIR¹, nous trouvons un lien avec les processus de maturation, de stress oxydatif et de stress au froid, pour 127 gènes (soit 61%). De plus, ces processus se retrouvent dans 88% des motifs de changement. Ces observations nous indiquent que la majorité des gènes formant un motif de changement peut être directement reliée à un processus biologique de la conservation du fruit à basse température après récolte.

De plus, nous n'avons pas d'information sur la fonction biologique pour 81 gènes du pommier qui ne sont donc pas considérés dans cette analyse. Néanmoins, les homologues d'*Arabidopsis* de certains d'entre eux sont cités dans des listes de gènes dérégulés dans différentes conditions de réponse à un stress, ce qui n'est donc pas inconsistant en soi.

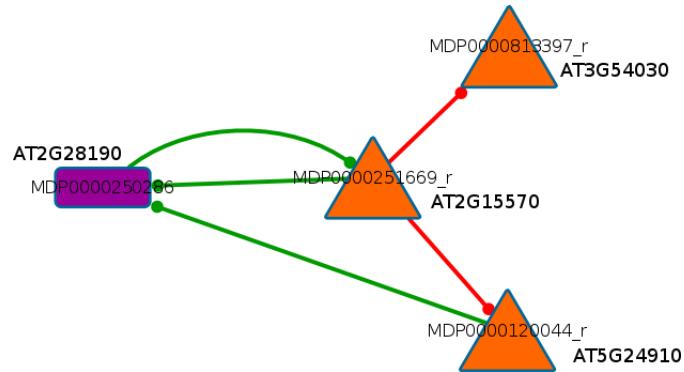
Nous pouvons expliciter un lien entre des motifs de changement et le processus de la maturation du fruit à basse température avec les trois exemples suivants.

Le motif de changement #1 (figure 6.8a) contient quatre gènes. MDP0000250286 code pour une superoxyde dismutase répondant à un stress de froid [Juszczak *et al.*, 2016]. MDP0000120044 est similaire à la cytochrome P450 monooxygénase CYP714A1 impliqué dans la voie de l'acide gibbérellique [Zhang *et al.*, 2011]. MDP0000251669 est similaire à une thiorédoxine [Issakidis-Bourguet *et al.*, 2001]. Ces trois gènes sont liés à une réponse à un stress oxydatif. Enfin, le quatrième gène MDP0000813397 code pour une kinase de la voie de signalisation des brassinostéroïdes [Tang *et al.*, 2008]. Les brassinostéroïdes sont des hormones de plantes impliquées dans la réponse à un stress froid [Clouse et Sasse, 1998]. Dans ce premier exemple, les acteurs du motif de changement sont tous directement liés à une réponse à un stress froid qui se produit lors du processus de stockage du fruit.

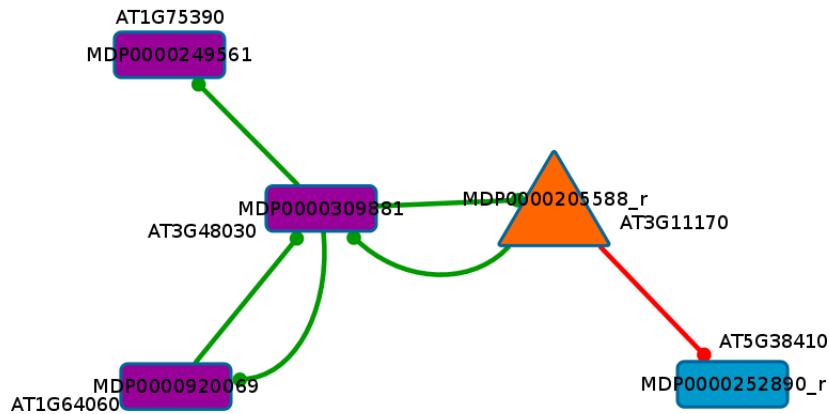
Le motif de changement #2 (figure 6.8b) contient deux gènes impliqués dans la réponse au froid : MDP0000205588, codant pour un acide gras désaturase [Shi *et al.*, 2011] et MDP0000249561, le facteur de transcription BZIP44 qui est également impliqué de manière putative dans le contrôle de la maturation du fruit *via* le rétrécissement de la paroi cellulaire [Iglesias-Fernández *et al.*, 2013, Weltmeier *et al.*, 2009]. Le motif contient également deux gènes impliqués dans la réponse à un stress oxydatif : MDP0000309881, est similaire au gène hypoxia-induced gene domain protein-1 (HIGD-1) [Hwang *et al.*, 2017] et MDP0000920069 code pour une protéine respiratory burst oxidase (RBO) F qui réagit également à l'éthylène et l'acide abscissique [Liu *et al.*, 2017]. Dans ce motif, seul le gène MDP0000252890 codant pour une sous-unité 3B de la Rubisco, ne peut pas être directement rattaché au processus étudié selon la littérature.

Sur les huit gènes du motif de changement #3 (figure 6.8c), cinq sont totalement cohérents avec le processus de la maturation du fruit à basse température. En effet, MDP0000250138 est un orthologue de BIN2, un membre de la famille d'ATSK (une kinase SHAGGY-like) qui agit dans la croisée des voies de signalisation de l'auxine et des brassinostéroïdes [Li *et al.*, 2017]. MDP0000797759, une protéine de liaison de l'ARN glycine-riche qui augmente la tolérance au stress sous des conditions de basse température [Kim *et al.*, 2010]. MDP0000151721 est un orthologue de Fro1 impliqué dans l'acclimatation au froid et

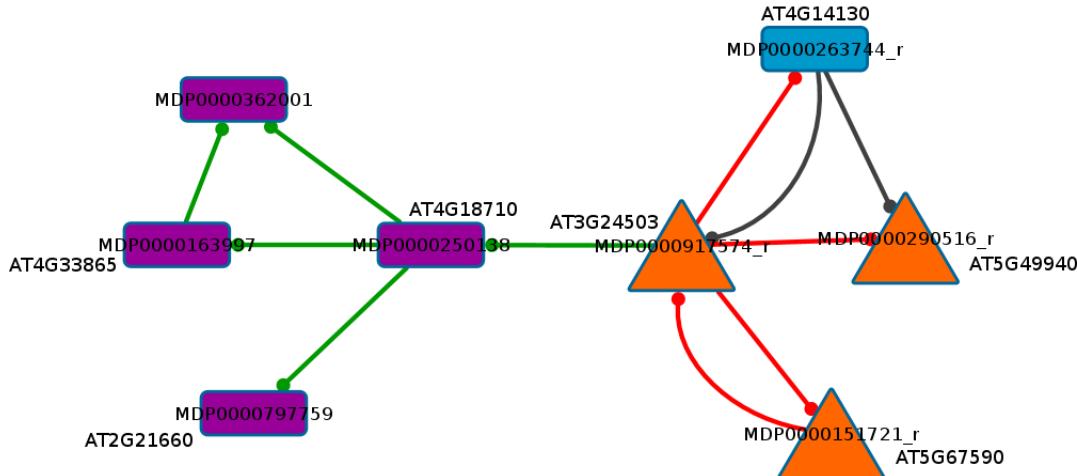
¹The *Arabidopsis* Information Resource (TAIR) - <https://www.arabidopsis.org/>



(a) Motif de changement #1. Le gène AS-impacté est MDP0000251669_r.



(b) Motif de changement #2. Le gène AS-impacté est MDP0000205588_r.



(c) Motif de changement #3. Le gène AS-impacté est MDP0000917574_r.

FIGURE 6.8 – Motifs de changement de l’expérience 60DAH. Chaque gène du pommier (MDP) est associé avec son meilleur homologue d’Arabidopsis (ATG) s’il existe.

la réponse au stress osmotique [Lee *et al.*, 2002]. MDP0000263744, un xyloglucane endotransglycosylase impliqué dans la modification de la paroi cellulaire et l’acclimatation au froid [Oono *et al.*, 2006]. Enfin MDP0000917574, un aldéhyde déshydrogénase 1A est un gène majeur dans la voie de signalisation des phénylpropanoïdes impliqués dans la production de composants antioxydants et dans la réponse aux stress biotiques et abiotiques [Nair *et al.*, 2004].

Nous montrons ainsi clairement que les fonctions biologiques de ces nouveaux acteurs sont liées à la question biologique étudiée. La considération des transcrits anti-sens permet de mettre en lumière de nouvelles régulations dans le réseau de gène.

De plus, une vue globale de l’ensemble des motifs de changement peut fournir un nouveau point de vue sur des voies de signalisation ou des gènes, dont l’importance aurait pu être sous-estimée sans l’intégration des nouvelles données. Dans notre cas, il est important de noter qu’au moins 11 occurrences de gènes liés à la voie de signalisation des brassinostéroïdes apparaissent parmi les 72 motifs de changement enrichis par une fonction de la GO slim. Les brassinostéroïdes sont impliqués dans divers processus du développement, tels que la croissance des racines et de la tige, l’initiation florale et le développement des fleurs et des fruits [Clouse et Sasse, 1998, Sasse, 2003]. Des études révèlent également que les brassinostéroïdes participent à la résistance des plantes contre différents stress abiotiques et biotiques, notamment celui contre le froid [Krishna, 2003, Bajguz et Hayat, 2009]. Dans l’étude de [Li *et al.*, 2012] il est également rapporté que le brassinolide régule la tolérance aux plantes contre les stress abiotiques en général et contre le froid en particulier. Les brassinostéroïdes sont également impliqués dans la maturation du raisin [Symons *et al.*, 2006] et dans le début du développement du fruit du concombre [Fu *et al.*, 2008], mais leur implication n’a pas encore été rapportée dans le développement du pommier et la maturation de la pomme. L’étude que nous avons menée montre que plusieurs transcrits sens et anti-sens sont impliqués dans la voie de signalisation des brassinostéroïdes et peuvent jouer un rôle dans la maturation de la pomme et dans la conservation du fruit à basse température.

Ces résultats montrent que l’intégration des données transcriptomiques anti-sens dans l’inférence de réseaux de gènes permet d’identifier, *via* les motifs de changement, de nouveaux acteurs reliés au processus biologique étudié.

6.3.3 Sous-graphes AS-impactés

L’analyse des motifs de changement et leur enrichissement fonctionnel nous donne des résultats intéressants et cohérents avec le contexte expérimental. Les motifs de changement offrent une vue locale de l’impact de l’intégration des données anti-sens dans l’inférence de réseau de gènes. Nous nous intéressons maintenant aux sous-graphes AS-impactés qui offrent une vue plus globale de cet impact en recherchant des modules dans le réseau de gènes.

Nous étudions les interactions dans le réseau SAS entre les gènes des sous-graphes AS-impactés. Un sous-graphe AS-impacté est composé uniquement de gènes AS-impactés, et, par définition, les interactions observées dans le sous-graphe AS-impacté ne sont plus directement représentées dans le cœur de

réseau SAS. Pour étudier cette reconfiguration, nous utilisons les arbres de Steiner.

Nous recherchons des sous-graphes AS-impactés contenant un minimum de trois nœuds. Dans l'expérience 60DAH nous identifions 29 sous-graphes AS-impactés qui contiennent au moins trois gènes. Pour chacun de ces sous-graphes AS-impactés, nous recherchons, dans le réseau SAS, un arbre de Steiner permettant de connecter l'ensemble des gènes AS-impactés constituant le sous-graphe d'origine. Nous avons utilisé l'heuristique des plus courts chemins. Peu de paires de gènes ont une distance ontologique qui diffère de la distance classique. Le temps de calcul supplémentaire pour la distance sémantique, et le faible gain d'information apporté par cette distance font que nous choisissons d'utiliser l'heuristique des plus courts chemins avec les distances classiques.

À partir des 29 sous-graphes AS-impactés, nous retrouvons 35 arbres de Steiner. La figure 6.9 représente un de ces arbres de Steiner. Cet arbre de Steiner connecte tous les gènes AS-impactés via les nœuds sens et anti-sens du réseau SAS. C'est un arbre composé de 26 nœuds terminaux (*i.e.* gènes AS-impactés) et de 82 nœuds de Steiner. On peut également observer sur la figure les interactions du sous-graphe AS-impacté. Avec cette visualisation, nous pouvons voir que le cœur de réseau a été grandement impacté par l'intégration des acteurs anti-sens, mais les gènes AS-impactés qui sont voisins directs dans le réseau S sont toujours connectés dans le réseau SAS de manière indirecte.

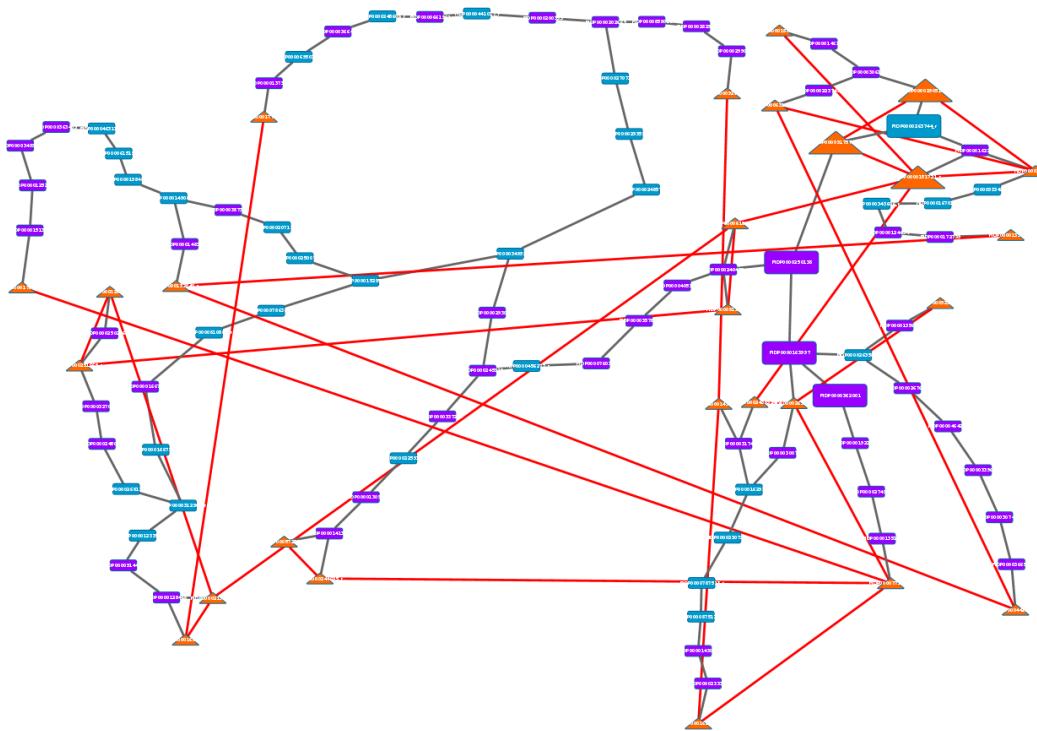


FIGURE 6.9 – Arbre de Steiner de la figure 5.9 dans lequel les nœuds du motif de changement #3 ont été mis en évidence. Les nœuds de plus grande taille sont ainsi les nœuds qui constituent le motif de changement #3.

L'arbre de Steiner de la figure 6.9 a la particularité de contenir un motif de changement. Le motif de changement #3 est en effet contenu dans cet arbre de Steiner. Nous nous intéressons donc aux fonctions

biologiques des gènes contenus dans l’arbre. Nous effectuons donc un enrichissement fonctionnel des gènes AS-impactés d’un côté et de l’ensemble des gènes de l’arbre de Steiner de l’autre. L’ensemble de gènes étant relativement petit par rapport à la taille du génome, nous utilisons la GO slim.

Sur cet arbre en particulier, nous obtenons deux termes de la GO slim qui enrichissent les gènes AS-impactés : “other metabolic processes” avec une p-valeur de 0.03001869, et “transcription, DNA-dependant” avec une p-valeur de 0.02386205. L’ensemble des 108 gènes de l’arbre de Steiner sont enrichis par un seul terme : “response to abiotic or biotic stimulus” avec une p-valeur de 0.04810813. L’enrichissement de l’ensemble des gènes de l’arbre est donc cohérent avec notre contexte expérimental.

Nous avons procédé à la même analyse sur l’ensemble des arbres de Steiner. Sur certains arbres on n’obtient pas d’enrichissement fonctionnel : l’ensemble des gènes est trop hétérogène pour être significatif pour un terme de la GO slim. Enfin on obtient parfois des catégories mais ces dernières ne permettent pas une interprétation biologique : “other metabolic processes” ou “unknown biological process” ne sont en effet pas des termes qui sont aisés à interpréter.

Conclusion et perspectives

Ce travail de thèse est une étude exploratoire de l'impact de l'intégration de données transcriptomiques anti-sens dans l'inférence de réseaux de gènes. Cette étude repose sur des principes d'analyse différentielle. Alors que les méthodes d'analyse différentielle s'effectuent sur un même ensemble d'acteurs dans deux conditions différentes, nous nous sommes intéressés à deux ensembles d'acteurs différents dans une même condition. Les deux ensembles d'acteurs que nous analysons sont les données de transcription sens d'un côté, et les données de transcription sens et anti-sens de l'autre.

La particularité de cette analyse transcriptomique est donc l'intégration de données anti-sens, dont les actions potentielles sur la régulation des gènes ont été assez peu étudiées et sont mal connues. Nos méthodes traitent les transcrits anti-sens de la même manière que les transcrits sens et les intègrent au jeu de données.

Notre étude s'appuie sur des données du pommier. Nous étudions une condition expérimentale en particulier, qui est une situation de stress pour le fruit : la maturation du fruit dans un environnement froid. En effet nous avons des données transcriptomiques sens et anti-sens de pommes au moment de la récolte, et deux mois après la récolte après qu'elles aient été stockées dans des chambres froides.

Nous définissons les *termes révélés par les anti-sens* dans notre analyse différentielle fonctionnelle. Ces termes sont des catégories de la Gene Ontology qui sont significativement sur-représentés dans les données sens et anti-sens, mais qui ne le sont pas dans le cas où les anti-sens ne sont pas utilisés. Nous identifions ainsi des processus biologiques en lien avec l'activité des transcrits anti-sens. Pour mener cette analyse fonctionnelle, nous avons dû proposer une annotation pour les transcrits anti-sens : nous avons associé le transcrit anti-sens à la même fonction biologique que le sens complémentaire. Cette hypothèse peut bien sûr être discutée, mais semble raisonnable en raison de la complémentarité de séquences entre le sens et l'anti-sens. De plus, l'interprétation biologique de nos résultats nous conduit à des processus biologiques en lien avec la condition expérimentale étudiée.

Avant de réaliser notre analyse différentielle de réseaux, nous proposons une méthode d'inférence de réseaux : *Extended Core Network (ECN)*. Cette méthode d'inférence, inspirée de C3NET, permet d'inférer un cœur de réseau constitué des meilleures interactions pour chacun des gènes. La méthode se base sur l'information mutuelle afin d'estimer la co-expression entre deux gènes. À l'aide d'un taux d'acceptation, la méthode identifie les meilleures interactions d'un gène.

Nous définissons une analyse différentielle de réseaux grâce à la méthode *Extended Core Network Best Interactions* qui fournit un cœur de réseau orienté permettant d'identifier les meilleures interactions

de chaque gène. Cette analyse différentielle de réseaux a la particularité de comparer deux réseaux issus de la même condition, mais inférés à partir de deux ensembles d'acteurs différents. La comparaison des réseaux révèle les *gènes AS-impactés* qui sont des gènes sens dont les meilleures interactions changent lorsqu'on intègre les données anti-sens. Un gène AS-impacté voit donc son voisinage dans le cœur de réseau totalement changé par l'intégration des transcrits anti-sens.

Un *motif de changement* rassemble autour d'un gène AS-impacté tous les acteurs concernés par l'impact des données anti-sens dans l'étude du réseau de gènes. L'interprétation biologique des motifs de changements indique que les anti-sens peuvent jouer un rôle dans la maturation du fruit lorsque celui-ci est stocké dans une chambre froide. En effet, nous observons, dans l'enrichissement fonctionnel des motifs de changements, plusieurs fonctions biologiques en lien avec la maturation du fruit et aussi avec la réponse au froid.

Cette thèse fait suite aux travaux menés sur les transcrits anti-sens au sein de l'IRHS [Celton *et al.*, 2014]. Ces travaux ont montré qu'un taux étonnamment élevé de transcrits anti-sens sont exprimés chez le pommier. Notre étude révèle un taux important de gènes AS-impactés et un lien fonctionnel avec les conditions d'expérimentation. Cela confirme l'intérêt d'étudier plus finement les transcrits anti-sens. Nos méthodes d'analyse permettent, à partir de données à large échelle, de fournir des ensembles restreints de gènes qui semblent fortement concernés par les effets de la transcription anti-sens, et qui méritent donc une validation expérimentale ciblée. Grâce aux méthodes développées, nous avons identifié plusieurs transcrits anti-sens qui peuvent jouer un rôle dans la maturation du fruit et dans la réponse au froid notamment. Il serait donc intéressant de concevoir des expériences biologiques spécifiques afin d'étudier le rôle de ces transcrits particuliers. Pour étudier des actions possibles des anti-sens, comme le Post Transcriptional Gene Silencing, les expériences de transcriptomique ne seront pas suffisantes ; il faudrait pouvoir également mesurer les quantités de protéines produites.

Les poursuites de cette thèse s'organisent en plusieurs axes ; d'une part, un axe plus biologique organisé autour de nouvelles connaissances sur le pommier, d'autre part, un axe plus bio-informatique autour de l'inférence de réseaux et de l'analyse différentielle de réseaux.

Un consortium international, monté et piloté par un membre de l'IRHS, vient d'obtenir un génome du pommier de très haute qualité, en combinant les dernières technologies de séquençage de l'ADN à celles de la cartographie classique [Daccord *et al.*, 2017]. L'IRHS est donc en train de concevoir une puce à ADN modélisée à partir de ce nouveau séquençage. Les méthodes que nous avons proposées pourront être reprises sur des données qui seront générées par cette nouvelle puce.

En attendant cela, les données disponibles que nous avons traitées peuvent être réutilisées pour une interprétation en fonction du nouveau génome. En effet, la puce AryANE v1.0 mesure les transcrits des gènes identifiés dans le génome de 2010, mais en fait, pour chaque sonde de la puce, on connaît la séquence d'ADN qu'elle représente. On peut donc rechercher une association entre une sonde de la

puce AryANE v1.0 et le nouveau génome du pommier. Dans cette correspondance, certaines sondes de la puce peuvent ne plus être associées à un gène du nouveau génome ; les informations correspondantes peuvent alors être retirées des jeux de données. Nous pouvons aussi avoir la situation où certains gènes du nouveau génome ne correspondent à aucune séquence de la puce AryANE v1.0. Pour eux, nous n'avons donc pas de mesures dans les données actuelles, ce qui est pénalisant pour une interprétation correcte de nos analyses. Il va nous falloir examiner les résultats que nous avons obtenus pour les mettre en correspondance avec le nouveau génome du pommier.

En plus du nouveau séquençage de l'ADN du pommier, les équipes de l'IRHS ont réalisé du séquençage orienté d'ARN. Cette méthode permet de mesurer directement l'expression d'un gène en séquençant les ARN transcrits, et de manière orientée afin de savoir s'il s'agit d'un transcrit sens ou anti-sens. Ces données RNA-Seq orientées ont l'avantage de séquencer l'ARN et donc il est plus facile d'identifier où débute et où termine la transcription de l'ARN, contrairement à la puce à ADN dont les sondes ont une taille limitée et donc ne correspondent qu'à une portion du transcrit. Cet avantage permet de connaître la séquence du transcrit, il est donc possible d'identifier quelles sont les premières bases du transcrit, et donc d'identifier plus précisément où se situe le site d'initiation de la transcription des anti-sens. Cela pourra être utile pour identifier les séquences promotrices des anti-sens afin de reprendre l'analyse que nous avons menée dans le chapitre 6.

Concernant la méthode d'analyse différentielle de réseaux que nous avons proposée, nous l'avons appliquée uniquement sur un cœur de réseau, constitué des interactions les plus fortes pour chaque gène. En effet, on sait que l'inférence de réseaux est un problème difficile et que les méthodes de reconstruction de réseaux obtiennent des résultats encore entachés de beaucoup d'erreurs. Lorsque l'on veut comparer deux réseaux inférés, l'incertitude est « en quelque sorte doublée » ; une interaction qui figure dans un réseau R_1 et ne se retrouve pas dans un réseau R_2 peut être un faux positif de l'inférence de R_1 ou un faux négatif de l'inférence de R_2 .

Pour effectuer l'analyse différentielle sur des réseaux de gènes, plus larges que les cœurs de réseaux, il faudrait donc avoir des réseaux constitués d'interactions robustes. Un moyen d'obtenir un noyau plus grand d'interactions robustes serait de construire un réseau consensus à partir des prédictions de plusieurs méthodes d'inférence. En effet, de nombreuses méthodes d'inférences, y compris celle que nous avons proposée, sont disponibles. Ces méthodes diffèrent sur la représentation du réseau, et également sur les méthodes d'identification d'une interaction : certaines méthodes estiment des relations paires à paires avec des mesures comme la corrélation ou l'information mutuelle, d'autres estiment les relations d'un gène avec tous les autres en utilisant des méthodes de régression ou d'apprentissage. Il serait intéressant de proposer une méthode combinant ces algorithmes (ensemble learning) afin de construire un noyau de réseau plus robuste, sur lequel les traitements que nous avons définis pourraient être utilisés.

L'analyse différentielle de réseaux est une discipline qui est de plus en plus utilisée dans différents travaux pour identifier des bio-marqueurs, des sous-réseaux, des ensembles de gènes, *etc.* C'est un domaine en plein essor, que nous avons pu appliquer ici dans un contexte particulier, où nous comparons

deux ensembles différents de gènes dans une même condition expérimentale. L'analyse différentielle, et particulièrement l'analyse différentielle de réseaux, est une discipline qui est peu utilisée dans le domaine végétal, mais qui a un potentiel pour identifier les spécificités d'une condition expérimentale, d'un mutant ou d'un génotype par exemple.

Bibliographie

- [Agrawal *et al.*, 1993] Agrawal, R., Imieliński, T., et Swami, A. (1993). Mining Association Rules Between Sets of Items in Large Databases. Dans *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, SIGMOD '93, pages 207–216, New York, NY, USA. ACM. (cité en page 50).
- [Altay *et al.*, 2011] Altay, G., Asim, M., Markowetz, F., et Neal, D. E. (2011). Differential C3net reveals disease networks of direct physical interactions. *BMC Bioinformatics*, 12(1) :296. (cité en page 56).
- [Altay et Emmert-Streib, 2010] Altay, G. et Emmert-Streib, F. (2010). Inferring the conservative causal core of gene regulatory networks. *BMC Systems Biology*, 4(1) :132. (cité en pages 41, 45 et 97).
- [Ashburner *et al.*, 2000] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., et Sherlock, G. (2000). Gene Ontology : tool for the unification of biology. *Nature Genetics*, 25(1) :25–29. (cité en page 80).
- [Auber, 2004] Auber, D. (2004). Tulip — A Huge Graph Visualization Framework. Dans Jünger, M. et Mutzel, P., éditeurs, *Graph Drawing Software*, Mathematics and Visualization, pages 105–126. Springer Berlin Heidelberg. (cité en page 38).
- [Bajguz et Hayat, 2009] Bajguz, A. et Hayat, S. (2009). Effects of brassinosteroids on the plant responses to environmental stresses. *Plant physiology and biochemistry : PPB*, 47(1) :1–8. (cité en page 139).
- [Bansal *et al.*, 2007] Bansal, M., Belcastro, V., Ambesi-Impiombato, A., et di Bernardo, D. (2007). How to infer gene networks from expression profiles. *Molecular Systems Biology*, 3(1) :78. (cité en page 41).
- [Barabási et Albert, 1999] Barabási, A.-L. et Albert, R. (1999). Emergence of Scaling in Random Networks. *Science*, 286(5439) :509–512. (cité en page 37).

- [Barabási *et al.*, 2011] Barabási, A.-L., Gulbahce, N., et Loscalzo, J. (2011). Network Medicine : A Network-based Approach to Human Disease. *Nature reviews. Genetics*, 12(1) :56–68. (cité en page 56).
- [Bastian *et al.*, 2009] Bastian, M., Heymann, S., et Jacomy, M. (2009). Gephi : An Open Source Software for Exploring and Manipulating Networks. Dans *Third International AAAI Conference on Weblogs and Social Media*. (cité en page 38).
- [Baulcombe, 2004] Baulcombe, D. (2004). RNA silencing in plants. *Nature*, 431(7006) :356–363. (cité en page 23).
- [Berretta et Morillon, 2009] Berretta, J. et Morillon, A. (2009). Pervasive transcription constitutes a new level of eukaryotic genome regulation. *EMBO Reports*, 10(9) :973–982. (cité en page 23).
- [Bockmayr *et al.*, 2013] Bockmayr, M., Klauschen, F., Györffy, B., Denkert, C., et Budczies, J. (2013). New network topology approaches reveal differential correlation patterns in breast cancer. *BMC Systems Biology*, 7 :78. (cité en page 56).
- [Borsani *et al.*, 2005] Borsani, O., Zhu, J., Verslues, P. E., Sunkar, R., et Zhu, J.-K. (2005). Endogenous siRNAs Derived from a Pair of Natural cis-Antisense Transcripts Regulate Salt Tolerance in Arabidopsis. *Cell*, 123(7) :1279–1291. (cité en page 62).
- [Brazhnik *et al.*, 2002] Brazhnik, P., Fuente, A. d. l., et Mendes, P. (2002). Gene networks : how to put the function in genomics. *Trends in Biotechnology*, 20(11) :467–472. (cité en pages 33 et 34).
- [Bulcke *et al.*, 2006] Bulcke, T. V. d., Leemput, K. V., Naudts, B., Remortel, P. v., Ma, H., Verschoren, A., Moor, B. D., et Marchal, K. (2006). SynTReN : a generator of synthetic gene expression data for design and analysis of structure learning algorithms. *BMC Bioinformatics*, 7(1) :43. (cité en pages 52 et 99).
- [Bullard *et al.*, 2010] Bullard, J. H., Purdom, E., Hansen, K. D., et Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, 11 :94. (cité en page 65).
- [Butte et Kohane, 2000] Butte, A. J. et Kohane, I. S. (2000). Mutual information relevance networks : functional genomic clustering using pairwise entropy measurements. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 418–429. (cité en pages 41 et 97).
- [Carbon *et al.*, 2009] Carbon, S., Ireland, A., Mungall, C. J., Shu, S., Marshall, B., et Lewis, S. (2009). AmiGO : online access to ontology and annotation data. *Bioinformatics*, 25(2) :288–289. (cité en page 84).

- [Celton *et al.*, 2014] Celton, J.-M., Gaillard, S., Bruneau, M., Pelletier, S., Aubourg, S., Martin-Magniette, M.-L., Navarro, L., Laurens, F., et Renou, J.-P. (2014). Widespread anti-sense transcription in apple is correlated with siRNA production and indicates a large potential for transcriptional and/or post-transcriptional control. *New Phytologist*, 203(1) :287–299. (cité en pages 14, 24, 60, 61, 62, 63, 64 et 144).
- [Chaparzadeh et Yavari, 2013] Chaparzadeh, N. et Yavari, B. (2013). Antioxidant responses of Golden delicious apple under cold storage conditions. *Iranian Journal of Plant Physiology*, 4(1) :907 – 915. (cité en page 119).
- [Chebil *et al.*, 2014] Chebil, I., Nicolle, R., Santini, G., Rouveiro, C., et Elati, M. (2014). Hybrid Method Inference for the Construction of Cooperative Regulatory Network in Human. *IEEE Transactions on NanoBioscience*, 13(2) :97–103. (cité en pages 41 et 50).
- [Chen *et al.*, 2006] Chen, X., Chen, M., et Ning, K. (2006). BNArray : an R package for constructing gene regulatory networks from microarray data by using Bayesian network. *Bioinformatics (Oxford, England)*, 22(23) :2952–2954. (cité en page 41).
- [Clouse et Sasse, 1998] Clouse, S. D. et Sasse, J. M. (1998). BRASSINOSTEROIDS : Essential Regulators of Plant Growth and Development. *Annual Review of Plant Physiology and Plant Molecular Biology*, 49 :427–451. (cité en pages 137 et 139).
- [Consortium, 2011] Consortium, A. I. M. (2011). Evidence for Network Evolution in an Arabidopsis Interactome Map. *Science*, 333(6042) :601–607. (cité en page 134).
- [Cover et Thomas, 1991] Cover, T. M. et Thomas, J. A. (1991). *Elements of Information Theory*. John Wiley & Sons. (cité en page 45).
- [Daccord *et al.*, 2017] Daccord, N., Celton, J.-M., Linsmith, G., Becker, C., Choisne, N., Schijlen, E., van de Geest, H., Bianco, L., Micheletti, D., Velasco, R., Di Pierro, E. A., Gouzy, J., Rees, D. J. G., Guérif, P., Muranty, H., Durel, C.-E., Laurens, F., Lespinasse, Y., Gaillard, S., Aubourg, S., Quesneville, H., Weigel, D., van de Weg, E., Troggio, M., et Bucher, E. (2017). High-quality de novo assembly of the apple genome and methylome dynamics of early fruit development. *Nature Genetics*, 49(7) :1099–1106. (cité en pages 60, 130, 136 et 144).
- [Davuluri *et al.*, 2003] Davuluri, R. V., Sun, H., Palaniswamy, S. K., Matthews, N., Molina, C., Kurtz, M., et Grotewold, E. (2003). AGRIS : Arabidopsis Gene Regulatory Information Server, an information resource of Arabidopsis cis-regulatory elements and transcription factors. *BMC Bioinformatics*, 4 :25. (cité en page 130).
- [de la Fuente, 2010] de la Fuente, A. (2010). From ‘differential expression’ to ‘differential networking’ – identification of dysfunctional regulatory networks in diseases. *Trends in Genetics*, 26(7) :326–333. (cité en page 54).

- [Dittrich *et al.*, 2008] Dittrich, M. T., Klau, G. W., Rosenwald, A., Dandekar, T., et Müller, T. (2008). Identifying functional modules in protein–protein interaction networks : an integrated exact approach. *Bioinformatics*, 24(13) :i223–i231. (cité en page 112).
- [Elati *et al.*, 2007] Elati, M., Neuvial, P., Bolotin-Fukuhara, M., Barillot, E., Radvanyi, F., et Rouveiro, C. (2007). LICORN : learning cooperative regulation networks from gene expression data. *Bioinformatics*, 23(18) :2407–2414. (cité en pages 41 et 50).
- [Emmert-Streib *et al.*, 2012] Emmert-Streib, F., Glazko, G., Gökmén, A., et De Matos Simoes, R. (2012). Statistical inference and reverse engineering of gene regulatory networks from observational expression data. *Bioinformatics and Computational Biology*, 3 :8. (cité en page 41).
- [English *et al.*, 1996] English, J. J., Mueller, E., et Baulcombe, D. C. (1996). Suppression of Virus Accumulation in Transgenic Plants Exhibiting Silencing of Nuclear Genes. *The Plant Cell*, 8(2) :179–188. (cité en page 23).
- [Faisal et Milenković, 2014] Faisal, F. E. et Milenković, T. (2014). Dynamic networks reveal key players in aging. *Bioinformatics (Oxford, England)*, 30(12) :1721–1729. (cité en page 55).
- [Faith *et al.*, 2007] Faith, J. J., Hayete, B., Thaden, J. T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J. J., et Gardner, T. S. (2007). Large-Scale Mapping and Validation of Escherichia coli Transcriptional Regulation from a Compendium of Expression Profiles. *PLOS Biology*, 5(1) :e8. (cité en page 41).
- [Fire *et al.*, 1998] Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E., et Mello, C. C. (1998). Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature*, 391(6669) :806–811. (cité en page 23).
- [Friedel *et al.*, 2012] Friedel, S., Usadel, B., von Wieren, N., et Sreenivasulu, N. (2012). Reverse Engineering : A Key Component of Systems Biology to Unravel Global Abiotic Stress Cross-Talk. *Frontiers in Plant Science*, 3 :294. (cité en page 41).
- [Fu *et al.*, 2008] Fu, F. Q., Mao, W. H., Shi, K., Zhou, Y. H., Asami, T., et Yu, J. Q. (2008). A role of brassinosteroids in early fruit development in cucumber. *Journal of Experimental Botany*, 59(9) :2299–2308. (cité en page 139).
- [Ge *et al.*, 2003] Ge, Y., Dudoit, S., et Speed, T. P. (2003). Resampling-based multiple testing for microarray data analysis. *Test*, 12(1) :1–77. (cité en page 83).
- [Gill *et al.*, 2014] Gill, R., Datta, S., et Datta, S. (2014). Differential network analysis in human cancer research. *Current Pharmaceutical Design*, 20(1) :4–10. (cité en page 54).

- [Golub *et al.*, 1999] Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., et Lander, E. S. (1999). Molecular classification of cancer : class discovery and class prediction by gene expression monitoring. *Science (New York, N.Y.)*, 286(5439) :531–537. (cité en page 54).
- [Haury *et al.*, 2012] Haury, A.-C., Mordelet, F., Vera-Licona, P., et Vert, J.-P. (2012). TIGRESS : Trustful Inference of Gene REgulation using Stability Selection. *BMC Systems Biology*, 6 :145. (cité en pages 41 et 48).
- [Horvath et Dong, 2008] Horvath, S. et Dong, J. (2008). Geometric Interpretation of Gene Coexpression Network Analysis. *PLoS Comput Biol*, 4(8) :e1000117. (cité en page 42).
- [Huynh-Thu *et al.*, 2010] Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., et Geurts, P. (2010). Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. *PLoS ONE*, 5(9) :e12776. (cité en pages 41 et 49).
- [Hwang *et al.*, 2017] Hwang, S.-T., Li, H., Alavilli, H., Lee, B.-H., et Choi, D. (2017). Molecular and physiological characterization of AtHIGD1 in Arabidopsis. *Biochemical and Biophysical Research Communications*, 487(4) :881–886. (cité en page 137).
- [Ideker et Krogan, 2012] Ideker, T. et Krogan, N. J. (2012). Differential network biology. *Molecular Systems Biology*, 8 :565. (cité en page 54).
- [Ideker *et al.*, 2002] Ideker, T., Ozier, O., Schwikowski, B., et Siegel, A. F. (2002). Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, 18(suppl_1) :S233–S240. (cité en page 55).
- [Iglesias-Fernández *et al.*, 2013] Iglesias-Fernández, R., Barrero-Sicilia, C., Carrillo-Barral, N., Oñate-Sánchez, L., et Carbonero, P. (2013). Arabidopsis thaliana bZIP44 : a transcription factor affecting seed germination and expression of the mannanase-encoding gene AtMAN7. *The Plant Journal : For Cell and Molecular Biology*, 74(5) :767–780. (cité en page 137).
- [Issakidis-Bourguet *et al.*, 2001] Issakidis-Bourguet, E., Mouaheb, N., Meyer, Y., et Miginiac-Maslow, M. (2001). Heterologous complementation of yeast reveals a new putative function for chloroplast m-type thioredoxin. *The Plant Journal : For Cell and Molecular Biology*, 25(2) :127–135. (cité en page 137).
- [Johnston *et al.*, 2002] Johnston, J. W., Hewett, E. W., et Hertog, M. L. A. T. M. (2002). Postharvest softening of apple (*Malus domestica*) fruit : A review. *New Zealand Journal of Crop and Horticultural Science*, 30(3) :145–160. (cité en page 119).
- [Juszczak *et al.*, 2016] Juszczak, I., Cvetkovic, J., Zuther, E., Hincha, D. K., et Baier, M. (2016). Natural Variation of Cold Deacclimation Correlates with Variation of Cold-Acclimation of the Plastid

- Antioxidant System in *Arabidopsis thaliana* Accessions. *Frontiers in Plant Science*, 7 :305. (cité en page 137).
- [Karp, 1972] Karp, R. M. (1972). Reducibility among Combinatorial Problems. Dans *Complexity of Computer Computations*, The IBM Research Symposia Series, pages 85–103. Springer, Boston, MA. (cité en page 111).
- [Kelley *et al.*, 2003] Kelley, B. P., Sharan, R., Karp, R. M., Sittler, T., Root, D. E., Stockwell, B. R., et Ideker, T. (2003). Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *Proceedings of the National Academy of Sciences of the United States of America*, 100(20) :11394–11399. (cité en page 54).
- [Kim *et al.*, 2010] Kim, J. Y., Kim, W. Y., Kwak, K. J., Oh, S. H., Han, Y. S., et Kang, H. (2010). Glycine-rich RNA-binding proteins are functionally conserved in *Arabidopsis thaliana* and *Oryza sativa* during cold adaptation process. *Journal of Experimental Botany*, 61(9) :2317–2325. (cité en page 137).
- [Krishna, 2003] Krishna, P. (2003). Brassinosteroid-Mediated Stress Responses. *Journal of Plant Growth Regulation*, 22(4) :289–297. (cité en page 139).
- [Kurt *et al.*, 2014] Kurt, Z., Aydin, N., et Altay, G. (2014). A Comprehensive Comparison of Association Estimators for Gene Network Inference Algorithms. *Bioinformatics*, page btu182. (cité en page 45).
- [Landeghem *et al.*, 2016] Landeghem, S. V., Parys, T. V., Dubois, M., Inzé, D., et de Peer, Y. V. (2016). Diffany : an ontology-driven framework to infer, visualise and analyse differential molecular networks. *BMC Bioinformatics*, 17 :18. (cité en page 56).
- [Langfelder et Horvath, 2008] Langfelder, P. et Horvath, S. (2008). WGCNA : an R package for weighted correlation network analysis. *BMC Bioinformatics*, 9(1) :559. (cité en pages 41, 42 et 57).
- [Langfelder *et al.*, 2011] Langfelder, P., Luo, R., Oldham, M. C., et Horvath, S. (2011). Is My Network Module Preserved and Reproducible ? *PLoS Computational Biology*, 7(1). (cité en pages 56 et 57).
- [Lee *et al.*, 2002] Lee, B.-h., Lee, H., Xiong, L., et Zhu, J.-K. (2002). A Mitochondrial Complex I Defect Impairs Cold-Regulated Nuclear Gene Expression. *The Plant Cell*, 14(6) :1235–1251. (cité en page 139).
- [Li *et al.*, 2012] Li, B., Zhang, C., Cao, B., Qin, G., Wang, W., et Tian, S. (2012). Brassinolide enhances cold stress tolerance of fruit by regulating plasma membrane proteins and lipids. *Amino Acids*, 43(6) :2469–2480. (cité en page 139).
- [Li *et al.*, 2017] Li, H., Ye, K., Shi, Y., Cheng, J., Zhang, X., et Yang, S. (2017). BZR1 Positively Regulates Freezing Tolerance via CBF-Dependent and CBF-Independent Pathways in *Arabidopsis*. *Molecular Plant*, 10(4) :545–559. (cité en page 137).

- [Liu *et al.*, 2017] Liu, B., Sun, L., Ma, L., et Hao, F.-S. (2017). Both AtrbohD and AtrbohF are essential for mediating responses to oxygen deficiency in Arabidopsis. *Plant Cell Reports*, 36(6) :947–957. (cité en page 137).
- [Luo *et al.*, 2013] Luo, C., Sidote, D. J., Zhang, Y., Kerstetter, R. A., Michael, T. P., et Lam, E. (2013). Integrative analysis of chromatin states in Arabidopsis identified potential regulatory mechanisms for natural antisense transcript production. *The Plant Journal : For Cell and Molecular Biology*, 73(1) :77–90. (cité en page 24).
- [Ma *et al.*, 2006] Ma, J., Morrow, D. J., Fernandes, J., et Walbot, V. (2006). Comparative profiling of the sense and antisense transcriptome of maize lines. *Genome Biology*, 7(3) :R22. (cité en page 61).
- [Maere *et al.*, 2005] Maere, S., Heymans, K., et Kuiper, M. (2005). BiNGO : a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks. *Bioinformatics*, 21(16) :3448–3449. (cité en pages 83 et 85).
- [Marbach *et al.*, 2012] Marbach, D., Costello, J. C., Küffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., Allison, K. R., The DREAM5 Consortium, Kellis, M., Collins, J. J., et Stolovitzky, G. (2012). Wisdom of crowds for robust gene network inference. *Nature Methods*, 9(8) :796–804. (cité en pages 41 et 42).
- [Marbach *et al.*, 2009] Marbach, D., Schaffter, T., Mattiussi, C., et Floreano, D. (2009). Generating realistic in silico gene networks for performance assessment of reverse engineering methods. *Journal of Computational Biology*, 16(2) :229–39. (cité en page 53).
- [Margolin *et al.*, 2006] Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R. D., et Califano, A. (2006). ARACNE : An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. *BMC Bioinformatics*, 7(Suppl 1) :S7. (cité en pages 41 et 45).
- [Maruyama *et al.*, 2012] Maruyama, R., Shipitsin, M., Choudhury, S., Wu, Z., Protopopov, A., Yao, J., Lo, P.-K., Bessarabova, M., Ishkin, A., Nikolsky, Y., Liu, X. S., Sukumar, S., et Polyak, K. (2012). Altered antisense-to-sense transcript ratios in breast cancer. *Proceedings of the National Academy of Sciences*, 109(8) :2820–2824. (cité en pages 24 et 63).
- [Medina-Rivera *et al.*, 2015] Medina-Rivera, A., Defrance, M., Sand, O., Herrmann, C., Castro-Mondragon, J. A., Delerce, J., Jaeger, S., Blanchet, C., Vincens, P., Caron, C., Staines, D. M., Contreras-Moreira, B., Artufel, M., Charbonnier-Khamvongsa, L., Hernandez, C., Thieffry, D., Thomas-Chollier, M., et van Helden, J. (2015). RSAT 2015 : Regulatory Sequence Analysis Tools. *Nucleic Acids Research*, 43(W1) :W50–W56. (cité en page 129).

- [Meyer *et al.*, 2007] Meyer, P. E., Kontos, K., Lafitte, F., et Bontempi, G. (2007). Information-Theoretic Inference of Large Transcriptional Regulatory Networks. *EURASIP Journal on Bioinformatics and Systems Biology*, 2007(1) :79879. (cité en pages 41 et 45).
- [Murray *et al.*, 2015] Murray, S. C., Haenni, S., Howe, F. S., Fischl, H., Chocian, K., Nair, A., et Mellor, J. (2015). Sense and antisense transcription are associated with distinct chromatin architectures across genes. *Nucleic Acids Research*, 43(16) :7823–7837. (cité en page 24).
- [Müssel *et al.*, 2010] Müssel, C., Hopfensitz, M., et Kestler, H. A. (2010). BoolNet—an R package for generation, reconstruction and analysis of Boolean networks. *Bioinformatics (Oxford, England)*, 26(10) :1378–1380. (cité en page 41).
- [Nair *et al.*, 2004] Nair, R. B., Bastress, K. L., Ruegger, M. O., Denault, J. W., et Chapple, C. (2004). The *Arabidopsis thaliana* REDUCED EPIDERMAL FLUORESCENCE1 gene encodes an aldehyde dehydrogenase involved in ferulic acid and sinapic acid biosynthesis. *The Plant Cell*, 16(2) :544–554. (cité en page 139).
- [Odibat et Reddy, 2012] Odibat, O. et Reddy, C. K. (2012). Ranking differential hubs in gene co-expression networks. *Journal of Bioinformatics and Computational Biology*, 10(1) :1240002. (cité en page 55).
- [Oono *et al.*, 2006] Oono, Y., Seki, M., Satou, M., Iida, K., Akiyama, K., Sakurai, T., Fujita, M., Yamaguchi-Shinozaki, K., et Shinozaki, K. (2006). Monitoring expression profiles of *Arabidopsis* genes during cold acclimation and deacclimation using DNA microarrays. *Functional & Integrative Genomics*, 6(3) :212–234. (cité en page 139).
- [Pelechano et Steinmetz, 2013] Pelechano, V. et Steinmetz, L. M. (2013). Gene regulation by antisense transcription. *Nature Reviews Genetics*, 14(12) :880–893. (cité en pages 24 et 25).
- [Perocchi *et al.*, 2007] Perocchi, F., Xu, Z., Clauder-Münster, S., et Steinmetz, L. M. (2007). Antisense artifacts in transcriptome microarray experiments are resolved by actinomycin D. *Nucleic Acids Research*, 35(19) :e128. (cité en page 65).
- [Prasanna *et al.*, 2007] Prasanna, V., Prabha, T. N., et Tharanathan, R. N. (2007). Fruit Ripening Phenomena—An Overview. *Critical Reviews in Food Science and Nutrition*, 47(1) :1–19. (cité en page 123).
- [Pržulj, 2007] Pržulj, N. (2007). Biological network comparison using graphlet degree distribution. *Bioinformatics*, 23(2) :e177–e183. (cité en page 37).
- [Qiu *et al.*, 2013] Qiu, X., Wu, H., et Hu, R. (2013). The impact of quantile and rank normalization procedures on the testing power of gene differential expression analysis. *BMC Bioinformatics*, 14 :124. (cité en page 65).

- [Reece *et al.*, 2011] Reece, J. B., Urry, L. A., Cain, M. L., Wasserman, S. A., Minorsky, P. V., Jackson, R. B., et others (2011). *Campbell biology*. Pearson Boston. (cité en page 18).
- [Sadeghi et Fröhlich, 2013] Sadeghi, A. et Fröhlich, H. (2013). Steiner tree methods for optimal sub-network identification : an empirical study. *BMC Bioinformatics*, 14 :144. (cité en pages 111 et 112).
- [Saito *et al.*, 2012] Saito, R., Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P.-L., Lotia, S., Pico, A. R., Bader, G. D., et Ideker, T. (2012). A travel guide to Cytoscape plugins. *Nature methods*, 9(11) :1069–1076. (cité en page 38).
- [Sasse, 2003] Sasse, J. M. (2003). Physiological Actions of Brassinosteroids : An Update. *Journal of Plant Growth Regulation*, 22(4) :276–288. (cité en page 139).
- [Schaffter *et al.*, 2011] Schaffter, T., Marbach, D., et Floreano, D. (2011). GeneNetWeaver : in silico benchmark generation and performance profiling of network inference methods. *Bioinformatics*, 27(16) :2263–2270. (cité en pages 52 et 53).
- [Shannon *et al.*, 2003] Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., et Ideker, T. (2003). Cytoscape : A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, 13(11) :2498–2504. (cité en page 38).
- [Sharan et Ideker, 2006] Sharan, R. et Ideker, T. (2006). Modeling cellular machinery through biological network comparison. *Nature Biotechnology*, 24(4) :427–433. (cité en page 54).
- [Sharp, 2001] Sharp, P. A. (2001). RNA interference—2001. *Genes & Development*, 15(5) :485–490. (cité en page 23).
- [Shi *et al.*, 2011] Shi, Y., An, L., Li, X., Huang, C., et Chen, G. (2011). The octadecanoid signaling pathway participates in the chilling-induced transcription of ω -3 fatty acid desaturases in Arabidopsis. *Plant physiology and biochemistry : PPB*, 49(2) :208–215. (cité en page 137).
- [Steuer *et al.*, 2002] Steuer, R., Kurths, J., Daub, C. O., Weise, J., et Selbig, J. (2002). The mutual information : Detecting and evaluating dependencies between variables. *Bioinformatics*, 18(suppl 2) :S231–S240. (cité en page 45).
- [Swiezewski *et al.*, 2009] Swiezewski, S., Liu, F., Magusin, A., et Dean, C. (2009). Cold-induced silencing by long antisense transcripts of an Arabidopsis Polycomb target. *Nature*, 462(7274) :799–802. (cité en pages 23 et 24).
- [Symons *et al.*, 2006] Symons, G. M., Davies, C., Shavrukov, Y., Dry, I. B., Reid, J. B., et Thomas, M. R. (2006). Grapes on Steroids. Brassinosteroids Are Involved in Grape Berry Ripening. *Plant Physiology*, 140(1) :150–158. (cité en page 139).

- [Tang *et al.*, 2008] Tang, W., Kim, T.-W., Oses-Prieto, J. A., Sun, Y., Deng, Z., Zhu, S., Wang, R., Burlingame, A. L., et Wang, Z.-Y. (2008). BSKs mediate signal transduction from the receptor kinase BRI1 in Arabidopsis. *Science (New York, N.Y.)*, 321(5888) :557–560. (cité en page 137).
- [Vanhée-Brossollet et Vaquero, 1998] Vanhée-Brossollet, C. et Vaquero, C. (1998). Do natural antisense transcripts make sense in eukaryotes ? *Gene*, 211(1) :1–9. (cité en page 61).
- [Velasco *et al.*, 2010] Velasco, R., Zharkikh, A., Affourtit, J., Dhingra, A., Cestaro, A., Kalyanaraman, A., Fontana, P., Bhatnagar, S. K., Troggio, M., Pruss, D., Salvi, S., Pindo, M., Baldi, P., Castelletti, S., Cavaiuolo, M., Coppola, G., Costa, F., Cova, V., Dal Ri, A., Goremykin, V., Komjanc, M., Longhi, S., Magnago, P., Malacarne, G., Malnoy, M., Micheletti, D., Moretto, M., Perazzolli, M., Si-Ammour, A., Vezzulli, S., Zini, E., Eldredge, G., Fitzgerald, L. M., Gutin, N., Lanchbury, J., Macalma, T., Mitchell, J. T., Reid, J., Wardell, B., Kodira, C., Chen, Z., Desany, B., Niazi, F., Palmer, M., Koepke, T., Jiwan, D., Schaeffer, S., Krishnan, V., Wu, C., Chu, V. T., King, S. T., Vick, J., Tao, Q., Mraz, A., Stormo, A., Stormo, K., Bogden, R., Ederle, D., Stella, A., Vecchietti, A., Kater, M. M., Masiero, S., Lasserre, P., Lespinasse, Y., Allan, A. C., Bus, V., Chagné, D., Crowhurst, R. N., Gleave, A. P., Lavezzo, E., Fawcett, J. A., Proost, S., Rouzé, P., Sterck, L., Toppo, S., Lazzari, B., Hellens, R. P., Durel, C.-E., Gutin, A., Bumgarner, R. E., Gardiner, S. E., Skolnick, M., Egholm, M., Van de Peer, Y., Salamini, F., et Viola, R. (2010). The genome of the domesticated apple (*Malus × domestica* Borkh.). *Nature Genetics*, 42(10) :833–839. (cité en pages 60 et 89).
- [Vincent *et al.*, 2015] Vincent, J., Martre, P., Gouriou, B., Ravel, C., Dai, Z., Petit, J.-M., et Pailloux, M. (2015). RuNet : A Web-Oriented Platform for Regulatory Network Inference, Application to Wheat –Omics Data. *PLOS ONE*, 10(5) :e0127127. (cité en page 51).
- [Wang *et al.*, 2014] Wang, H., Chung, P. J., Liu, J., Jang, I.-C., Kean, M. J., Xu, J., et Chua, N.-H. (2014). Genome-wide identification of long noncoding natural antisense transcripts and their responses to light in arabidopsis. *Genome research*, 24(3) :444–453. (cité en page 24).
- [Wang *et al.*, 2005] Wang, X.-J., Gaasterland, T., et Chua, N.-H. (2005). Genome-wide prediction and identification of cis-natural antisense transcripts in *Arabidopsis thaliana*. *Genome Biology*, 6(4) :R30. (cité en pages 24, 61, 62 et 63).
- [Weltmeier *et al.*, 2009] Weltmeier, F., Rahmani, F., Ehlert, A., Dietrich, K., Schütze, K., Wang, X., Chaban, C., Hanson, J., Teige, M., Harter, K., Vicente-Carbajosa, J., Smeekens, S., et Dröge-Laser, W. (2009). Expression patterns within the *Arabidopsis* C/S1 bZIP transcription factor network : availability of heterodimerization partners controls gene expression during stress response and development. *Plant Molecular Biology*, 69(1-2) :107–119. (cité en page 137).
- [Yamaguchi-Shinozaki et Shinozaki, 2005] Yamaguchi-Shinozaki, K. et Shinozaki, K. (2005). Organization of cis-acting regulatory elements in osmotic- and cold-stress-responsive promoters. *Trends in Plant Science*, 10(2) :88–94. (cité en page 129).

- [Yang *et al.*, 2012] Yang, Q., Rao, J., Yi, S., Meng, K., Wu, J., et Hou, Y. (2012). Antioxidant enzyme activity and chilling injury during low-temperature storage of Kiwifruit cv. Hongyang exposed to gradual postharvest cooling. *Horticulture, Environment, and Biotechnology*, 53(6) :505–512. (cité en page 119).
- [Yang *et al.*, 2002] Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J., et Speed, T. P. (2002). Normalization for cDNA microarray data : a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research*, 30(4) :e15. (cité en page 53).
- [Yu *et al.*, 2010] Yu, G., Li, F., Qin, Y., Bo, X., Wu, Y., et Wang, S. (2010). GOSemSim : an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics*, 26(7) :976–978. (cité en page 115).
- [Zhang et Horvath, 2005] Zhang, B. et Horvath, S. (2005). A General Framework for Weighted Gene Co-Expression Network Analysis. *Statistical Applications in Genetics and Molecular Biology*, 4(1). (cité en page 42).
- [Zhang *et al.*, 2013] Zhang, X., Liu, K., Liu, Z.-P., Duval, B., Richer, J.-M., Zhao, X.-M., Hao, J.-K., et Chen, L. (2013). NARROMI : a noise and redundancy reduction technique improves accuracy of gene regulatory network inference. *Bioinformatics*, 29(1) :106–113. (cité en pages 41 et 48).
- [Zhang *et al.*, 2011] Zhang, Y., Zhang, B., Yan, D., Dong, W., Yang, W., Li, Q., Zeng, L., Wang, J., Wang, L., Hicks, L. M., et He, Z. (2011). Two Arabidopsis cytochrome P450 monooxygenases, CYP714a1 and CYP714a2, function redundantly in plant development through gibberellin deactivation. *The Plant Journal : For Cell and Molecular Biology*, 67(2) :342–353. (cité en page 137).



Liste des motifs de changements enrichis de 60DAH

La table liste les termes issus de l'enrichissement fonctionnel de chacun des motifs de changements de l'expérience 60DAH. Les motifs sont déterminés par les quatre groupes d'acteurs différents :

- **AS-impacté** : le gène AS-impacté du motif
- **voisins S** : voisins du gène AS-impacté dans le réseau S
- **voisins SAS** : voisins du gène AS-impacté dans le réseau SAS
- **voisins AS** : voisins des anti-sens dans le réseau SAS

Pour chaque motif, la **catégorie GO** et la **p-valeur** associée sont indiquées. Parmi les 304 motifs de changement, seuls les 72 qui ont un enrichissement sont ici listés.

AS-impacté	voisins S	voisins SAS	voisins AS	catégorie GO	p-valeur
MDP0000119630_r	MDP0000347520_r	MDP0000193127 MDP0000198054	MDP0000573331 MDP0000937996	other biological processes	0.0384
MDP0000126738_r	MDP0000614645_r	MDP0000269319	MDP0000126738_r	developmental processes	0.0386
MDP0000129456_r	MDP0000794936_r	MDP0000753318	MDP0000319081 MDP0000129456_r MDP0000842877_r	cell organization and biogenesis	0.0077
MDP0000133105_r	MDP0000131368_r	MDP0000288678	MDP0000228370	other metabolic processes	0.0290

AS-impacté	voisins S	voisins SAS	voisins AS	catégorie GO	p-valeur
MDP0000136801_r	MDP0000227978_r	MDP0000347813	MDP0000345210_r	unknown biological processes	0.0242
MDP0000137211_r	MDP0000197472_r MDP0000199588_r MDP0000236041_r MDP0000277710_r	MDP0000189485 MDP0000221813 MDP0000252854	MDP0000308890 MDP0000137211_r MDP0000277710_r MDP0000277710_r	other metabolic processes	0.0302
MDP0000139736_r	MDP0000293592_r	MDP0000807958	MDP0000293592_r	response to abiotic or biotic stimulus	0.0487
MDP0000146108_r	MDP0000302279_r	MDP0000216918	MDP0000146108_r	transcription, DNA-dependent	0.0191
MDP0000152278_r	Md_miR167c_r MDP0000217451_r MDP0000231210_r	MDP0000686661	MDP0000253300_r	other biological processes	0.0384
MDP0000158544_r	MDP0000295794_r	MDP0000158544	MDP0000158544_r	cell organization and biogenesis	0.0260
MDP0000159873_r	MDP0000196762_r	MDP0000564897	MDP0000159873_r	response to abiotic or biotic stimulus	0.0487
MDP0000175504_r	MDP0000427102_r MDP0000464562_r	MDP0000146698	MDP0000135168	other biological processes	0.0214
MDP0000178329_r	MDP0000123085_r	MDP0000168388	MDP0000286450 MDP0000314479 MDP0000398010 MDP0000318341_r	developmental processes	0.0400
MDP0000185543_r	MDP0000120231_r MDP0000126738_r MDP0000231845_r MDP0000430546_r MDP0000507003_r MDP0000613689_r	MDP0000839889	MDP0000159873_r MDP0000185543_r	response to stress	0.0478
MDP0000189326_r	MDP0000232609_r	MDP0000126194	MDP0000189326_r	developmental processes	0.0386
MDP0000191837_r	MDP0000225179_r	MDP0000409061	MDP0000582937	cell organization and biogenesis	0.0488
MDP0000194076_r	MDP0000214562_r MDP0000881155_r	MDP0000745371	MDP0000254055_r	other biological processes	0.0214
MDP0000196174_r	MDP0000199159_r	MDP0000358501	MDP0000263613	unknown biological processes	0.0242
MDP0000196373_r	MDP0000147181_r	MDP0000760376	MDP0000324254	response to abiotic or biotic stimulus	0.0085

AS-impacté	voisins S	voisins SAS	voisins AS	catégorie GO	p-valeur
MDP0000196762_r	MDP0000159873_r	MDP0000564897	MDP0000159873_r	response to abiotic or biotic stimulus	0.0487
MDP0000205588_r	MDP0000252890_r	MDP0000309881	MDP0000249561 MDP0000920069 MDP0000205588_r	other metabolic processes	0.0120
MDP0000216726_r	MDP0000863909_r	MDP0000195837	MDP0000169619	other metabolic processes	0.0290
MDP0000218122_r	MDP0000230627_r	MDP0000349802	MDP0000218122_r	electron transport or energy pathways	0.0460
MDP0000221908_r	MDP0000210504_r MDP0000230511_r	MDP0000171728 MDP0000662344	MDP0000186977 MDP0000187714 MDP0000301534	transport	0.0287
MDP0000225179_r	MDP0000191837_r MDP0000263744_r	MDP0000582937	MDP0000225179_r	other cellular processes	0.0336
MDP0000225939_r	MDP0000155229_r	MDP0000130950 MDP0000337205	MDP0000141372 MDP0000245918 MDP0000225939_r	other metabolic processes	0.0472
MDP0000228370_r	MDP0000173207_r MDP0000940098_r	MDP0000456476	MDP0000228370_r	other metabolic processes	0.0290
MDP0000231674_r	MDP0000268112_r MDP0000307286_r	MDP0000313525	MDP0000465844 MDP0000499668 MDP0000813710	transport	0.0192
MDP0000231960_r	MDP0000199442_r MDP0000220633_r	MDP0000921319	MDP0000231960_r	transport	0.0427
MDP0000233098_r	MDP0000302532_r	MDP0000398010	MDP0000547069 MDP0000320534_r	response to stress	0.0353
MDP0000240073_r	MDP0000938969_r	MDP0000307441	MDP0000240073_r	response to abiotic or biotic stimulus	0.0487
MDP0000243895_r	MDP0000151721_r MDP0000263161_r	MDP0000317457	MDP0000289630 MDP0000615656	other metabolic processes	0.0472
MDP0000247926_r	MDP0000119446_r MDP0000225179_r MDP0000347092_r	MDP0000363724	MDP0000052249	developmental processes	0.0250
MDP0000251669_r	MDP0000120044_r MDP0000813397_r	MDP0000250286	MDP0000251669_r	other metabolic processes	0.0290
MDP0000257492_r	MDP0000671114_r	MDP0000831519	MDP0000195614	response to abiotic or biotic stimulus	0.0085

AS-impacté	voisins S	voisins SAS	voisins AS	catégorie GO	p-valeur
MDP0000264060_r	MDP0000191887_r MDP0000200853_r MDP0000449818_r MDP0000598183_r MDP0000709066_r MDP0000930401_r	MDP0000190553	MDP0000264060_r	response to abiotic or biotic stimulus	0.0142
MDP0000264639_r	MDP0000244867_r MDP0000681768_r	MDP0000222512	MDP0000264639_r	transport	0.0427
MDP0000267323_r	MDP0000200696_r	MDP0000234570	MDP0000267323_r	transport	0.0227
MDP0000274441_r	MDP0000119148_r	MDP0000231959	MDP0000326734	cell organization and biogenesis	0.0488
MDP0000275654_r	MDP0000210504_r MDP0000319355_r MDP0000834570_r	MDP0000343606	MDP0000291221_r	transport	0.0117
MDP0000285243_r	MDP0000507003_r	MDP0000232389	MDP0000285243_r	response to abiotic or biotic stimulus	0.0487
MDP0000287083_r	MDP0000200696_r MDP0000236041_r	MDP0000125336 MDP0000223735	MDP0000261874 MDP0000940098	transport	0.0192
MDP0000287992_r	MDP0000206586_r MDP0000302532_r	MDP0000303061	MDP0000287992_r	developmental processes	0.0060
MDP0000290516_r	MDP0000306141_r	MDP0000306185	MDP0000146108	cell organization and biogenesis	0.0488
MDP0000292031_r	MDP0000933677_r	MDP0000196802	MDP0000292031_r	response to abiotic or biotic stimulus	0.0487
MDP0000300808_r	MDP0000868045_r	MDP0000231748	MDP0000251669_r	response to stress	0.0161
MDP0000304482_r	MDP0000251531_r MDP0000254055_r MDP0000449818_r	MDP0000304620	MDP0000889811	other biological processes	0.0043
MDP0000306141_r	MDP0000933422_r	MDP0000162244	MDP0000151721_r	cell organization and biogenesis	0.0488
MDP0000312744_r	MDP0000257492_r MDP0000530206_r	MDP0000632173	MDP0000312744_r	signal transduction	0.0387
MDP0000314478_r	MDP0000270657_r	MDP0000314479	MDP0000160965	unknown biological processes	0.0242
MDP0000326493_r	MDP0000210504_r MDP0000230511_r	MDP0000163005	MDP0000326493_r	transport	0.0427
MDP0000349837_r	MDP0000138851_r MDP0000167665_r MDP0000196593_r MDP0000229959_r MDP0000269868_r MDP0000908370_r	MDP0000242083	MDP0000349837_r	unknown biological processes	0.0464

AS-impacté	voisins S	voisins SAS	voisins AS	catégorie GO	p-valeur
MDP0000356275_r	MDP0000150068_r MDP0000911376_r	MDP0000312765	MDP0000222512 MDP0000239321_r MDP0000356275_r	other metabolic processes	0.0472
MDP0000358379_r	MDP0000759336_r	MDP0000288293	MDP0000206314 MDP0000268229 MDP0000360414	response to abiotic or biotic stimulus	0.0348
MDP0000362615_r	MDP0000301534_r	MDP0000362615	MDP0000362615_r	transport	0.0007
MDP0000363455_r	MDP0000938969_r	MDP0000250138 MDP0000318172 MDP0000356405	MDP0000163997 MDP0000318172 MDP0000362001 MDP0000797759 MDP0000797759	transport	0.0287
MDP0000369382_r	MDP0000233098_r MDP0000235446_r MDP0000464276_r	MDP0000398010	MDP0000547069 MDP0000320534_r	response to stress	0.0175
MDP0000371774_r	MDP0000142895_r	MDP0000148501	MDP0000142895_r	other biological processes	0.0026
MDP0000419926_r	MDP0000158470_r	MDP0000419926	MDP0000158470_r MDP0000419926_r	response to abiotic or biotic stimulus	0.0023
MDP0000452739_r	MDP0000173207_r	MDP0000090281 MDP0000456476	MDP0000173207_r MDP0000228370_r MDP0000228370_r MDP0000452739_r MDP0000940098_r	other metabolic processes	0.0472
MDP0000484244_r	MDP0000474349_r	MDP0000272650	MDP0000340236_r	unknown biological processes	0.0242
MDP0000534716_r	MDP0000193127_r MDP0000698024_r	MDP0000534716	MDP0000534716_r	transport	0.0427
MDP0000572721_r	MDP0000244605_r	MDP0000141372	MDP0000244605_r MDP0000572721_r	transport	0.0227
MDP0000628096_r	MDP0000239321_r MDP0000270835_r	MDP0000725398	MDP0000440796	protein metabolism	0.0257
MDP0000658332_r	MDP0000244605_r	MDP0000561026	MDP0000658332_r	transport	0.0227
MDP0000694624_r	MDP0000220633_r MDP0000547461_r	MDP0000760376	MDP0000324254	other biological processes	0.0214
MDP0000745680_r	MDP0000251783_r	MDP0000132480	MDP0000504159	protein metabolism	0.0115
MDP0000813397_r	MDP0000618088_r	MDP0000240478 MDP0000250138	MDP0000163997 MDP0000250138 MDP0000362001 MDP0000797759 MDP0000618088_r	protein metabolism	0.0115

AS-impacté	voisins S	voisins SAS	voisins AS	catégorie GO	p-valeur
MDP0000835914_r	MDP0000283947_r	MDP0000187320	MDP0000835914_r	electron transport or energy pathways	0.0460
MDP0000917574_r	MDP000151721_r MDP000263744_r MDP000290516_r	MDP000250138	MDP000163997 MDP000362001 MDP000797759	other metabolic processes	0.0104
MDP0000923601_r	MDP000194795_r	MDP000604171	MDP000323490	unknown biological processes	0.0242
MDP0000941318_r	MDP000295562_r	MDP000858039	MDP000202669 MDP000314385 MDP000930401_r MDP000941318_r	response to abiotic or biotic stimulus	0.0348



B

Étude des données à partir de WGCNA

Comme nous l'avons signalé dans le chapitre 3, nous avons utilisé le logiciel WGCNA afin de réaliser une étude exploratoire sur les données pommier. Lors de cette étude nous avons inféré un réseau de gènes et analysé sa topologie grâce aux modules de la méthode WGCNA.

Nous avons construit un réseau de gène pour l'expérience 60DAH. Nous introduisons ce réseau dans le chapitre 3 et la répartition des gènes dans les modules du réseau de gènes est rappelée dans la table B.1. Nous avons d'abord réalisé un enrichissement fonctionnel sur les gènes du modules, enfin nous avons construit un réseau pour l'expérience H également et nous comparons les modules de ces deux réseaux.

Enrichissement fonctionnel

Nous avons réalisé un enrichissement fonctionnel des modules de 60DAH. Sur le même principe que l'analyse fonctionnelle différentielle proposée dans le chapitre 4, nous avons réalisé l'enrichissement fonctionnel d'un module avec uniquement ses transcrits sens d'un côté, et avec l'ensemble des transcrits sens et anti-sens de l'autre. L'enrichissement fonctionnel a été réalisé sur la GO slim que nous avons créée pour les données pommier. Seuls les termes dont la p-valeur est inférieure à 0.05 sont gardés à l'issue de l'enrichissement fonctionnel. Les résultats de ces deux enrichissements fonctionnels sont montrés dans la table B.2.

La première observation que l'on peut faire est que, parmi les gènes qui n'ont pas été affectés à un module (le module « grey »), la réponse à un stimulus est une fonction biologique qui est significativement sur-représentée dans les deux analyses. Lorsque l'analyse est réalisée avec les transcrits sens et anti-sens, on note en plus les fonctions liées au transport et à une réponse au stress. WGCNA met donc de côté des gènes qui jouent peut-être un rôle dans la maturation du fruit.

TABLE B.1 – Répartition des transcrits sens et anti-sens dans les modules du réseau 60DAH. Il est indiqué la taille des modules, le nombre de transcrits sens (# S) et anti-sens (# AS) contenus dans chacun module.

60DAH				
Module	Taille module	# S	# AS	
grey	669	418	251	
turquoise	313	169	144	
brown	95	28	67	
red	50	31	19	
greenyellow	24	17	7	
green	72	27	45	
yellow	85	50	35	
blue	133	87	46	
pink	41	24	17	
magenta	27	17	10	
purple	26	11	15	
salmon	21	7	14	
black	46	26	20	
tan	23	19	4	

La deuxième observation que nous faisons est que cinq modules n'ont pas d'enrichissement fonctionnel en n'utilisant que les données sens, mais un enrichissement est trouvé lorsque toutes les données sens et anti-sens sont utilisées. Le module « blue » notamment, composé de 87 transcrits sens et 46 transcrits anti-sens, n'a donc aucun terme fonctionnel sur-représenté par les transcrits sens uniquement, mais trois termes différents lorsque les transcrits sens et anti-sens sont utilisés. Parmi ces trois termes on peut noter la présence de “response to abiotic or biotic stimulus” et de “response to stress”, qui sont deux termes liés à la condition d’expérimentation.

La troisième observation porte sur le module « greenyellow ». Lorsque seuls les transcrits sens sont utilisés pour l'enrichissement fonctionnel, on obtient le terme “transcription,DNA-dependent”, mais lorsque les transcrits sens et anti-sens sont utilisés, on obtient le terme “transport”. Cela peut s'expliquer par le fait que le module « greenyellow » est une module très petit ne contenant que 17 transcrits sens et 7 transcrits anti-sens. Les sept transcrits anti-sens n'étant pas étiquetés par le terme “transcription,DNA-dependent”, la p-valeur associée à ce terme a donc augmenté et il ne se retrouve plus sur-représenté. À l'inverse, les transcrits anti-sens étant associés au terme “transport”, sa p-valeur associée a diminuée et passe ainsi le seuil de 0.05.

Comparaison entre H et 60DAH

WGCNA est un logiciel qui permet de comparer des modules de réseaux entre eux. Nous l'avons donc utilisé afin d'inférer un réseau pour H et ensuite comparer les modules du réseau H avec les modules du réseau 60DAH. Nous souhaitons ainsi savoir si les gènes d'un module de H se retrouvent dans un même

TABLE B.2 – Termes de l'enrichissement fonctionnel des transcrits des modules de 60DAH. Nous indiquons les termes de la GO slim, que nous avons construite pour les données pommier, pour les transcrits sens et pour les transcrits sens et anti-sens qui composent le module.

Module	Enrichissement sens	Enrichissement sens et anti-sens
grey	response to abiotic or biotic stimulus	response to abiotic or biotic stimulus transport response to stress
turquoise	-	transport
brown	-	response to abiotic or biotic stimulus
red	electron transport or energy pathways	electron transport or energy pathways
greenyellow	transcription,DNA-dependent	transport
green	-	-
yellow	-	response to abiotic or biotic stimulus
blue	-	other biological processes response to abiotic or biotic stimulus response to stress
pink	-	response to abiotic or biotic stimulus
magenta	electron transport or energy pathways	electron transport or energy pathways
purple	cell organization and biogenesis	cell organization and biogenesis
salmon	-	-
black	-	-
tan	other metabolic processes electron transport or energy pathways	other metabolic processes electron transport or energy pathways

module de 60DAH. Pour cela nous avons utilisé l'outil de correspondance des modules. Cet outil indique le nombre de transcrits communs entre les deux modules comparés. Il indique également la conservation du module, c'est-à-dire si l'intersection des deux modules est représentative pour les deux modules.

La table B.3 indique la répartition des gènes dans les modules du réseau H. La figure B.1 montre les correspondances des modules entre H et 60DAH. Les modules de H sont indiqués en ordonnée, avec le nombre total de transcrits compris dans le module. Les modules de 60DAH sont indiqués en abscisse, également avec le nombre total de transcrits dans chaque module. Les nombres dans la matrice représentent le nombre de transcrits identiques entre les modules de H et 60DAH. Le dégradé de rouge d'une case de la matrice représente la conservation du module : plus la case est rouge plus le module est préservé.

En observant la figure B.1, on observe que certains modules de l'expérience H se retrouvent dans l'expérience 60DAH. Afin de savoir si la consistance de ces modules est due plus par les transcrits sens ou anti-sens, nous avons fait la même correspondance mais qu'avec les transcrits sens d'un côté (figure B.1 en bas à gauche) et qu'avec les transcrits anti-sens de l'autre (figure B.1 en bas à droite).

On observe alors que le module gris est très conservé pour les anti-sens. Cela signifie que les mêmes transcrits anti-sens n'ont pas été assignés à un module dans H et 60DAH. On n'observe pas la même chose avec les transcrits sens, ce qui signifie que les transcrits sens sont plus affectés par le changement de condition entre H et 60DAH que les transcrits anti-sens.

TABLE B.3 – Termes de l'enrichissement fonctionnel des transcrits des modules de H. Nous indiquons les termes de la GO slim, que nous avons construite pour les données pommier, pour les transcrits sens et pour les transcrits sens et anti-sens qui composent le module.

Module	Taille module	# S	# AS
grey	820	545	275
turquoise	256	88	168
brown	56	40	16
blue	151	89	62
green	52	18	34
black	45	21	24
tan	23	9	14
red	47	20	27
purple	29	15	14
pink	35	19	16
yellow	54	25	29
greenyellow	27	21	6
magenta	30	21	9

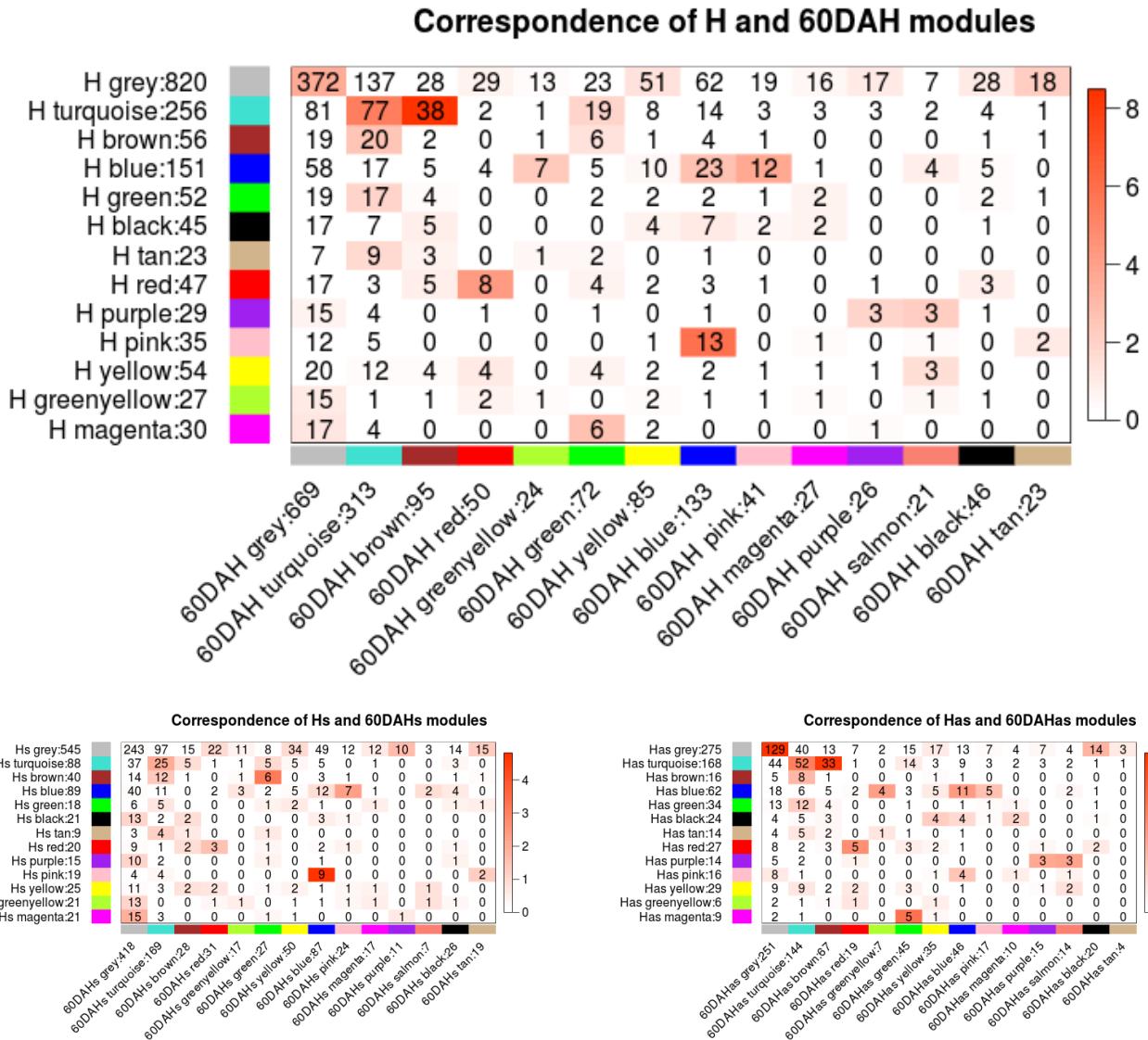


FIGURE B.1 – Correspondance des réseaux H et 60DAH avec l'ensemble des transcrits (en haut), avec seulement les transcrits sens (à gauche) et avec seulement les transcrits anti-sens (à droite).



Liste de mes contributions

Article de revue internationale en cours de soumission

Suite à la publication dans la conférence internationale IWBBIO 2017 de l'article « Differential network analysis of anti-sense regulation », il nous a été proposé d'étendre ce papier pour publication dans le journal BMC Systems Biology. L'article a été envoyé le 1^{er} septembre 2017 et une réponse est attendue en fin d'année.

Conférences internationales

1. Marc Legeay, Béatrice Duval, et Jean-Pierre Renou. Differential network analysis of anti-sense regulation. Dans Ignacio Rojas et Francisco M. Ortúñoz Guzman, éditeurs, *Bioinformatics and Biomedical Engineering - 5th International Work-Conference, IWBBIO 2017, Granada, Spain, April 26-28, 2017, Proceedings, Part II*, volume 10209 de *Lecture Notes in Computer Science*, pages 277–288, 2017.
2. Marc Legeay, Béatrice Duval, et Jean-Pierre Renou. Inference and differential analysis of extended core networks : A way to study anti-sense regulation. Dans Tianhai Tian, Qinghua Jiang, Yunlong Liu, Kevin Burrage, Jiangning Song, Yadong Wang, Xiaohua Hu, Shinichi Morishita, Qian Zhu, et Guohua Wang, éditeurs, *IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2016, Shenzhen, China, December 15-18, 2016*, pages 284–287. IEEE Computer Society, 2016.
3. Marc Legeay, Béatrice Duval, et Jean-Pierre Renou. Differential functional analysis and change motifs in gene networks to explore the role of anti-sense transcription. Dans Anu G. Bourgeois,

Pavel Skums, Xiang Wan, et Alex Zelikovsky, éditeurs, *Bioinformatics Research and Applications - 12th International Symposium, ISBRA 2016, Minsk, Belarus, June 5-8, 2016, Proceedings*, volume 9683 de *Lecture Notes in Computer Science*, pages 117–126. Springer, 2016.

Communications internationales

4. Marc Legeay, Béatrice Duval, et Jean-Pierre Renou. Context-specific co-expression networks to explore the impact of anti-sense transcription. Dans *Poster aux Computational Methods in Systems Biology (CMSB)*, 2015.

Communications nationales

5. Marc Legeay et Béatrice Duval. Comparaison de réseaux de gènes pour explorer le rôle des transcrits anti-sens. Dans *Grands Graphes et Bioinformatique, atelier d'Extraction et Gestion des Connaissances (EGC)*, 2016.
6. Marc Legeay. Construction et analyse de réseaux contextuels de co-expression pour des données transcriptomiques sens et anti-sens. Dans *Communication orale au Réseau méthodologique MIA "Inférence de réseaux (biologiques)" (NETBIO)*, 2015.
7. Marc Legeay, Béatrice Duval, Jean-Pierre Renou, et Julie Bourbeillon. Construction et analyse de réseaux de gènes contextuels dans le domaine végétal. Dans *Poster aux Journées Ouvertes en Biologie, Informatique & Mathématiques (JOBIM)*, 2015.

Publications suite au travail de Master

Mon travail de Master portait sur l’interrogation des graphes conceptuels. Nous nous sommes basés sur le langage d’interrogation du Web sémantique, SPARQL, pour créer un nouveau langage d’interrogation.

8. David Genest, Marc Legeay, Stéphane Loiseau, et Christophe Bechade. A graphical language to query conceptual graphs. Dans *26th IEEE International Conference on Tools with Artificial Intelligence, ICTAI 2014, Limassol, Cyprus, November 10-12, 2014*, pages 304–308. IEEE Computer Society, 2014.
9. Marc Legeay, David Genest, et Stéphane Loiseau. Un langage d’interrogation à la SPARQL pour les graphes conceptuels. Dans Benoît Otjacques, Jérôme Darmont, and Thomas Tamisier, editors, *15èmes Journées Francophones Extraction et Gestion des Connaissances, EGC 2015, 27-30 Janvier 2015, Luxembourg*, volume E-28 de *Revue des Nouvelles Technologies de l’Information*, pages 227–238. Hermann-Éditions, 2015.

Thèse de Doctorat

Marc LEGEAY

Étude de la régulation anti-sens par l'analyse différentielle de données transcriptomiques dans le domaine végétal

Study of the anti-sense regulation by differential analysis of transcriptomic data in plants

Résumé

Un des problèmes actuels en bio-informatique est de comprendre les mécanismes de régulation au sein d'une cellule ou d'un organisme. L'objectif de la thèse est d'étudier les réseaux de co-expression de gènes chez le pommier avec la particularité d'y intégrer les transcrits anti-sens. Les transcrits anti-sens sont des ARN généralement non-codants, dont les différents modes d'action sont encore mal connus. Dans notre étude exploratoire du rôle des anti-sens, nous proposons d'une part une analyse fonctionnelle différentielle qui met en évidence l'intérêt de l'intégration des données anti-sens en transcriptomique. D'autre part, concernant les réseaux de gènes, nous proposons de limiter l'inférence à un cœur de réseau et nous introduisons alors une méthode d'analyse différentielle permettant de comparer un réseau obtenu à partir de données sens avec un réseau contenant des données sens et anti-sens. Nous introduisons ainsi la notion de gènes AS-impacté, qui permet d'identifier des gènes dont les interactions au sein d'un réseau de co-expression sont fortement impactées par la prise en compte de transcrits anti-sens. Pour les données pommier que nous avons étudiées et qui concerne la maturation des fruits et leur conservation à basse température, l'interprétation biologique des résultats de notre analyse différentielle fournit des pistes pertinentes pour une étude expérimentale plus ciblée de gènes ou de voies de signalisation dont l'importance pourrait être sous-estimée sans la prise en compte des données anti-sens.

Mots clés

bio-informatique, fouille de données, réseaux de gènes, analyse différentielle, transcription anti-sens.

Abstract

A challenging task in bioinformatics is to decipher cell regulation mechanisms. The objective of this thesis is to study gene networks from apple data with the particularity to integrate anti-sense transcription data. Anti-sense transcripts are mostly non-coding RNAs and their different roles in the cell are still not well known. In our study, to explore the role of anti-sense transcripts, we first propose a differential functional analysis that highlights the interest of integrating anti-sense data into a transcriptomic analysis. Then, regarding gene networks, we propose to focus on inference of a core network and we introduce a new differential analysis method that allows to compare a sense network with a sense and anti-sense network. We thus introduce the notion of AS-impacted genes, that allows to identify genes that are highly co-expressed with anti-sense transcripts. We analysed apple data related to ripening of fruits stored in cold storage; biological interpretation of the results of our differential analysis provides some promising leads to a more targeted experimental study of genes or pathways, which role could be underestimated without integration of anti-sense data.

Key Words

bioinformatics, data mining, gene networks, differential analysis, anti-sense transcription.