Modélisation d'un réseau de régulation d'ARN pour prédire des fonctions de gènes impliqués dans le mode de reproduction du puceron du pois

Valentin Wucher

ANNÉE 2014

UNIVERSITÉ DE RENNES 1

THÈSE / UNIVERSITÉ DE RENNES 1
sous le sceau de l'Université Européenne de Bretagne

pour le grade de

DOCTEUR DE L'UNIVERSITÉ DE RENNES 1

Mention : Biologie

École doctorale Vie-Agro-Santé

présentée par

Valentin WUCHER

préparée à l'unité mixte de recherche UMR1349 – IGEPP et
UMR6074 – IRISA
Institut de Génétique, Environnement et Protection des Plantes et
Institut de Recherche en Informatique et Système Aléatoires
Composante Universitaire SVE

Modélisation d'un réseau de régulation d'ARN pour prédire des fonctions de gènes impliqués dans le mode de reproduction du puceron du pois

Thèse soutenue à Rennes
le 3 Novembre 2014

devant le jury composé de :

Christian DIOT
Directeur de recherche INRA / *Président*
Roderic GUIGO
Research director CRG Barcelone / *Rapporteur*
Emmanuelle JACQUIN-JOLY
Directrice de recherche INRA / *Rapporteure*
Julien BOBE
Directeur de recherche INRA / *Examinateur*
Hélène TOUZET
Directrice de recherche CNRS / *Examinatrice*
Denis TAGU
Directeur de recherche INRA / *Directeur de thèse*

Remerciements

Je tiens tout d'abord à remercier mes deux directeurs : Jacques et Denis, ou Denis et Jacques (pour ne pas faire de jaloux). Merci beaucoup de m'avoir donné la possibilité de faire cette thèse avec vous, d'avoir offert sa chance au petit gars de la capitale. Merci de l'aide que vous m'avez apportée sur tous les aspects professionnels de ma vie de thésard, d'avoir pris le temps de me transmettre vos connaissances et de m'avoir montré ce qu'était le métier de chercheur. Ça a été un vrai plaisir de passer ces trois années sous votre direction à tous les deux. Merci.

Merci à l'ensemble de mon jury de thèse : Roderic Guigo, Emmanuelle Jacquin-Joly, Julien Bode, Christian Diot et Hélène Touzet pour avoir accepté d'évaluer mon travail.

Merci aussi à mon comité de thèse : Yann Audic, Hervé Seitz, Hélène Touzet, Emmanuelle Becker, Gaël Le Trionnaire, Fabrice Legeai et Anne Siegel. Merci pour les discussions enrichissantes que nous avons eues pendant mes deux comités et pour vos conseils.

Je tiens aussi à remercier deux personnes qui m'ont aidé pendant mon parcours d'étudiant : Onnik Agbulut et Stéphane Le Crom, deux de mes directeurs de stages précédents. Merci à vous deux pour vos encouragements, je ne serai sûrement pas là aujourd'hui sans votre aide et votre soutien.

Le travail serait différent sans de bons collègues de bureau : merci Cyril pour les moments de détentes et les franches rigolades. Merci aussi à mon ancien collègue de bureau, Thomas, qui le premier m'a fait découvrir cette ville de Rennes. Je remercie aussi tous les amis et collègues que j'ai pu rencontrer pendant ces trois années au sein de mes deux équipes, Pierre (les deux), Clovis, Nicolas, Guillaume, Renaud, Claire, Nathalie, Raluca, Sylvain, Gaël, Lucie et j'en oublie sûrement. Un remerciement spécial pour Aurore avec qui j'ai pu parler, sans tabous, du puceron. De même pour Fabrice, que j'ai embêté maintes et maintes fois.

Je finis en remerciant toutes les personnes extérieures au monde de la recherche qui m'ont soutenu ; la famille : môm, Philippe, Adrien, Claire, Patrick. Je remercie aussi celle que j'ai rencontrée à la fin de ma thèse et qui m'a supporté jusqu'à son aboutissement : Adriana. Merci à tous les amis de Paris : Benjamin, Axel, Romain, Zoé, Amandine, Anne-Sophie, Jean-Octave, Olivier, Héloïse, Nabil, Yohan et bien d'autres.

Préambule

Dans ce manuscrit, sont présentés les résultats en français et synthétisés de mes travaux de thèse. J'ai participé à la rédaction d'un article de synthèse sur la plasticité du mode de reproduction chez les pucerons. J'ai également publié un article scientifique suite à la sélection d'un résumé pour une conférence lors du congrès European Conference on Data Analysis 2013. Ces deux articles sont présentés en Annexe. Un article est en cours de rédaction sur les ARN non-codant chez *Acyrthosiphon pisum* et leur régulation lors du changement de mode de reproduction :

Wucher V., Legeai F., Jaubert-Possamai S., le Trionnaire G., Hudaverdian S., Nicolas J., Seitz H., Siegel A., Derrien T., Gallot A. and Tagu D. : Non-coding RNAs profiling : molecular network involved in phenotypic plasticity of the pea aphid.

Il agrège une grande partie de mes résultats sur les microARN et les réseaux d'interactions ainsi que d'autres travaux sur les ARN piwi et les longs ARN non codants auxquels j'ai également participé.

Table des matières

Chapitre 1

Introduction

1.1 Introduction générale : étudier un réseau de régulation impliqué dans le polyphénisme de reproduction chez le puceron

1.1.1 Les réseaux : une représentation des différentes régulations biologiques et un moyen d'intégrer des données hétérogènes

L'arrivée à la fois de la génération de données à haut débit en biologie et de la bioinformatique a permis l'émergence d'une nouvelle discipline : la biologie des systèmes. Cette discipline consiste à étudier à grande échelle les génomes, les transcriptomes, les protéomes ou encore les métabolomes, pour ne citer qu'eux, et à intégrer, mettre en relation ces différents niveaux d'information du fonctionnement du génome. Cela a permis de mettre en évidence des relations existantes entre différentes molécules au sein d'une même catégorie de molécules mais aussi entre ces différentes catégories. Ces approches intégratives à grande échelle permettent d'avoir une vision systémique d'un processus biologique avec un ensemble d'acteurs agissant en synergie plutôt que comme différents acteurs agissant de manière plus ou moins isolée. Néanmoins, devant le nombre très élevé d'éléments et de relations, des outils spécifiques doivent être mis au point pour visualiser, interpréter les différentes relations existantes ainsi qu'en déduire de nouvelles connaissances. Il s'agit par exemple d'étudier les relations entre les 1.800 facteurs de transcription et les 20.000 gènes humains ou encore les interactions protéines-protéines possibles entre les 100.000 protéines humaines. La modélisation par des réseaux de ces relations et leur étude permet de tendre vers ce but.

Les réseaux, qui sont mathématiquement des graphes, sont un moyen de représentation, de visualisation et d'analyse de données très utile dans toutes les applications informatiques. L'analyse des réseaux permet d'extraire de la connaissance issue d'interactions complexes entre objets et est un outil d'aide à la décision : la lecture et l'interprétation des réseaux permettent de poser des hypothèses de travail sur le processus étudié. Chaque élément est représenté par un nœud dans le graphe et chaque relation entre deux éléments est représentée par une arête, orientée ou non, entre deux nœuds. En biologie moléculaire, on trouve notamment des réseaux métaboliques, des réseaux d'interactions protéines-protéines ou encore des réseaux de régulation transcriptionnelle par des facteurs de transcriptions.

Avec l'émergence en biologie des relations à grande échelle décrites plus haut, l'analyse des réseaux prend une importance croissante. En analysant la topologie de ces graphes, c'est-à-dire la façon dont sont répartis les nœuds et les arêtes, on peut mettre en évidence des groupes spécifiques de relations ou des éléments clés pour formuler de nouvelles hypothèses et aider à sélectionner des expérimentations biologiques pouvant permettre de formuler de nouvelles conclusions sur le phénomène étudié. Par exemple, la Figure 1.1 montre un réseau de gènes impliqués dans des maladies où chaque nœud représente un gène avec une couleur qui indique la type de maladie et une arête signifie que les deux gènes sont impliqués dans un même maladie. Cette visualisation permet de mettre en évidence des regroupements de gènes suivant le nombre d'arêtes qu'ils ont entre eux et que ces regroupements correspondent aux couleurs/maladies associées aux gènes.

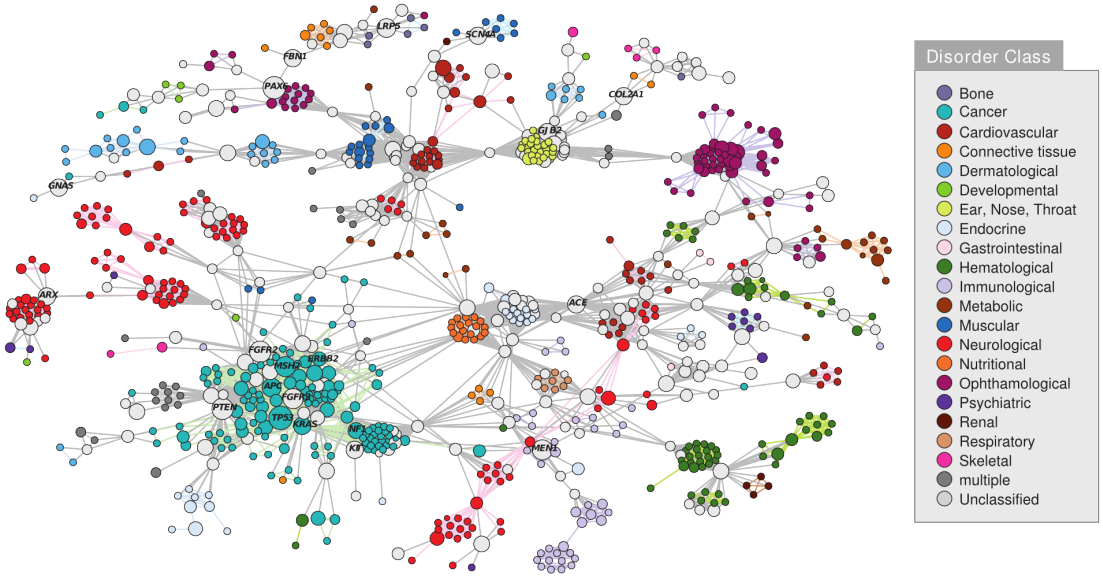


FIG. 1.1 – Réseau de gènes impliqués dans des pathologies humaines. Chaque nœud représente un gène et il existe une arête entre deux gènes s'ils sont impliqués dans une même maladie. Image tirée de Goh *et al.* [1].

1.1.2 La plasticité phénotypique : un des modes d'adaptation des espèces aux changements de leur environnement

Classiquement en biologie, le dogme veut que l'on associe à un génome un phénotype. Nous pouvons voir ça de la façon suivante : à un ensemble d'informations (le génome et son ensemble de gènes), on associe une lecture de ces informations, le phénotype. La plasticité phénotypique au sens large peut-être définie comme la capacité d'un organisme à répondre à un facteur environnemental en modifiant son phénotype au niveau physiologique, biochimique ou comportemental. Pour reprendre la vision précédente, la plasticité phénotypique peut être vue comme la capacité d'un organisme à lire l'information contenue dans son génome de façon différente selon le contexte ou les différentes conditions expérimentales.

La plasticité phénotypique est présente chez de nombreux organismes que ce soit dans des bactéries comme *Escherichia coli* qui répond différemment à la présence de lactose dans son environnement [2], chez les plantes avec l'acacia *Acacia drepanolobium* qui répond à l'attaque d'herbivores par la production de tiges avec de grosses épines [3] ou encore chez les insectes avec la taille des ailes de *Drosophila melanogaster* qui varie en fonction de la température ou la chenille *Nemoria arizonaria* qui mime son habitat [4] (Figure 1.2). Parmi les différents cas de plasticité phénotypique, on peut distinguer deux classes : continue et discrète [5]. La plasticité phénotypique continue est caractérisée par un nombre élevé de phénotypes qui évoluent de façon graduelle entre deux phénotypes extrêmes comme le cas des ailes *D. melanogaster* dont la taille varie de façon continue en fonction du nombre de cellules qui les constituent en passant par toutes les valeurs des plus petites aux plus grandes. La plasticité phénotypique discrète, appelée polyphénisme, se dit pour un organisme lorsque que pour son génome

les phénotypes alternatifs sont très différenciés et sans intermédiaire. Il existe différents types de polyphénisme. On peut notamment citer chez les insectes le polyphénisme de caste où les individus n'ont pas la même physiologie ni le même rôle au sein de la communauté comme chez les fourmis ou les abeilles ; le polyphénisme de dispersion avec des individus ailés ou non comme chez les pucerons ; le polyphénisme de reproduction avec des individus possédant différents modes de reproductions, présent encore une fois chez les pucerons.

Le polyphénisme constitue un trait biologique très particulier où les facteurs environnementaux influencent le développement des individus en modifiant ou ré-orientant des programmes génétiques et par conséquent les différentes régulations opérant sur le génome, le transcriptome, le protéome, etc.

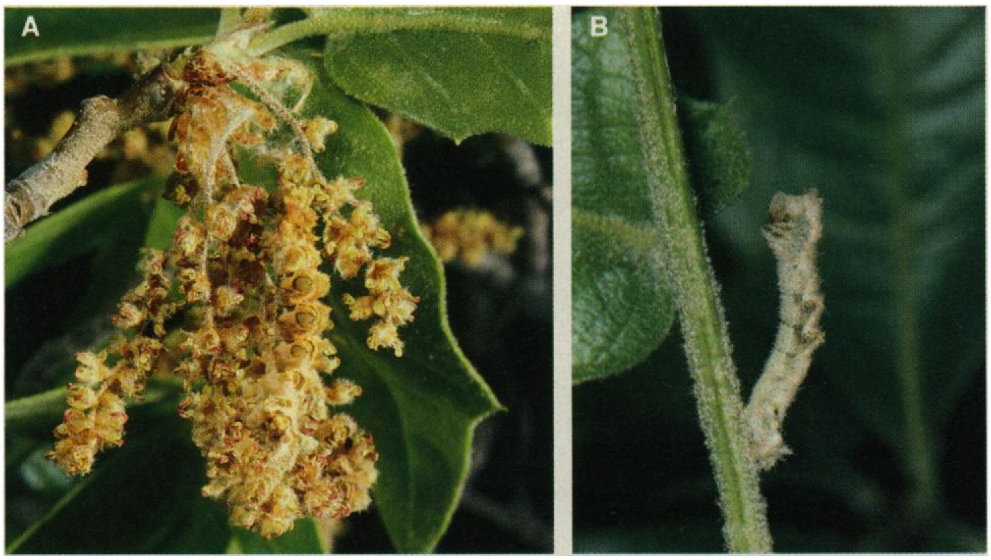


FIG. 1.2 – Image des deux phénotypes de la chenille *Nemoria arizonaria*. À gauche une chenille nourrie avec des fleurs de chênes, ressemblant à ces fleurs. À droite, une chenille nourrie avec des feuilles de chênes, ressemblant à une brindille. Image tirée de Greene [4].

Plan de l'introduction

La plasticité phénotypique du mode de reproduction est étudiée dans l'équipe INRA dans laquelle j'ai travaillé. Afin d'appréhender la complexité des mécanismes mis en œuvres, l'équipe développe une approche intégrative, via des réseaux de gènes [6].

Lors de cette thèse, le réseau de régulation post-transcriptionnelle entre microARN et ARNm a été étudié dans le contexte de la plasticité phénotypique de reproduction chez *Acyrthosiphon pisum*, le puceron du pois, à l'aide d'une méthode d'analyse de relations binaires, l'analyse de concepts formels. Ainsi les paragraphes suivants décrivent :

1. Les microARN ;
2. Le polyphénisme de reproduction du puceron du pois ;
3. L'analyse de concepts formels.

1.2 Régulation post-transcriptionnelle par les petits ARN non codant : le cas des microARN

Pendant longtemps, on a considéré que les régions du génome qui étaient transcrites mais non traduites, c'est-à-dire des régions qui ne codent pas pour des protéines, étaient de l'ADN inutile et que ces transcrits étaient dégradés par la cellule. Cette idée a été remise en cause avec la découverte de nouveaux types de régulation par des ARN non codants (ARNnc). En effet, certains ARNnc ont été identifiés comme étant impliqués dans des régulations épigénétiques, transcriptionnelles ou encore post-transcriptionnelles. La régulation de la stabilité et de la traduction des ARNm par les microARN fait partie de ces régulations post-transcriptionnelles par les ARNnc.

1.2.1 Les microARN : découverte et identification

Découverte des microARN chez les animaux

La première description d'un microARN remonte à 1993 avec la découverte chez *Caenorhabditis elegans* d'un petit ARN issus du gène lin-4 qui est complémentaire de séquences répétées dans le 3'UTR de l'ARNm du gène lin-14 [7, 8, 9]. Lee *et al.* et Wightman *et al.* observent que lin-4 est nécessaire à la transition entre les stades L1 et L2 du développement larvaire [7, 8]. L'inactivation de lin-14 produit un phénotype opposé à la perte de lin-4 ; les auteurs en déduisent que lin-4 régule négativement l'expression de lin-14 [9]. Plus tard en 2000, Reinhart *et al.* montrent qu'un ARN de 21 nucléotides, let-7, régule le développement chez *C. elegans* [10] dont la séquence a été trouvée comme étant conservée chez plusieurs espèces notamment chez *Drosophila melanogaster* et l'humain [11]. La dénomination « microARN » a été donnée en 2001 dans trois publications par Lagos-Quintana *et al.*, Lau *et al.* et Lee et Ambros [12, 13, 14]. Aujourd'hui, on sait qu'un microARN mature (fonctionnel) se présente généralement comme un ARN simple-brin d'environ 22 nucléotides. La chaîne de réaction permettant de passer d'un gène de microARN à un microARN mature est schématisée Figure 1.3. Ce microARN mature est issu de la transcription de son gène en un pri-microARN, long ARN. Ce pri-microARN possède une ou plusieurs structures en tige-boucle qui sont ensuite clivées pour former des pré-microARN. La boucle d'un pré-microARN est ensuite clivée pour former un ARN double-brin à partir duquel le microARN mature est libéré et intégré à un complexe nucléoprotéique appelé RISC. La structure en tige-boucle des différents précurseurs est très conservée car elle est nécessaire à la reconnaissance et la maturation de ces précurseurs. La biosynthèse des microARN est décrite plus en détail ci-dessous partie 1.2.3. La fonction d'inhibition des microARN matures est principalement portée par ce que l'on appelle la *graine* du microARN mature, classiquement définie comme les six nucléotides situés entre les positions 2 et 7 incluses, qui est complémentaire d'une zone de son/ses ARNm cible(s), préférentiellement située à l'extrémité 3'UTR de l'ARNm. Cette reconnaissance aboutit dans certains cas à une diminution ou un arrêt de la traduction de la protéine de l'ARNm reconnu et à une baisse de sa stabilité. Les microARN sont donc des régulateurs post-transcriptionnels des gènes.

En 2009, dans sa revue sur les microARN [16], Bartel stipule qu'il peut s'avérer difficile de trouver des fonctions biologiques qui ne sont pas influencées par les microARN. Depuis cette date, l'implication des microARN chez de nombreuses espèces et dans de

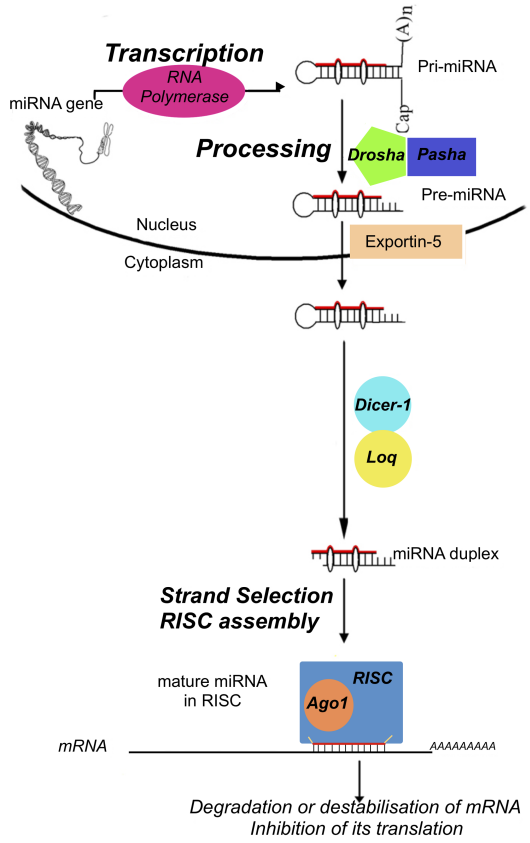


FIG. 1.3 – Voie de biosynthèse des microARN. Figure adaptée de IAGC [15].

nombreux processus biologiques a été mise en évidence. Chez les insectes, des fonctions comme le vieillissement, la régulation de l'insuline, la croissance cellulaire ou encore la reproduction sont soumis au contrôle de microARN [17]. Les microARN ont aussi un rôle dans de nombreuses maladies. miR2Disease est une base de données qui regroupe un ensemble de microARN impliqué dans différentes maladies [18].

Identification des microARN par méthodes haut débit

Depuis la découverte des premiers microARN lin-4 et let-7 (qui à cette époque n'étaient pas encore appelés « microARN »), de nombreux microARN ont été identifiés grâce au développement du séquençage à haut débit. La base de données miRBase [19, 20, 21, 22, 23] regroupe actuellement plus de 28.000 microARN pour 223 espèces différentes (pour la version 21 de juin 2014 de miRBase). Par exemple, le génome humain contient 2.588 microARN matures, 1.915 pour la souris, 434 pour *C. elegans* ou encore 427 pour *Arabidopsis thaliana*. La Figure 1.4 présente le nombre de microARN présent dans miRBase ainsi que le nombre de publications contenant le mot « microRNA » dans PubMed.

Les microARN étant codés par les génomes, plusieurs méthodes existent pour la prédiction des gènes de microARN, soit *in silico* en utilisant le génome, soit en couplant ces prédictions avec des données de séquençage de petits ARN. Parmi les critères utilisés

par ces méthodes on peut citer la stabilité et la taille de la structure en tige-boucle correspondant au pré-microARN ou la conservation des gènes de microARN chez des espèces proches [24].

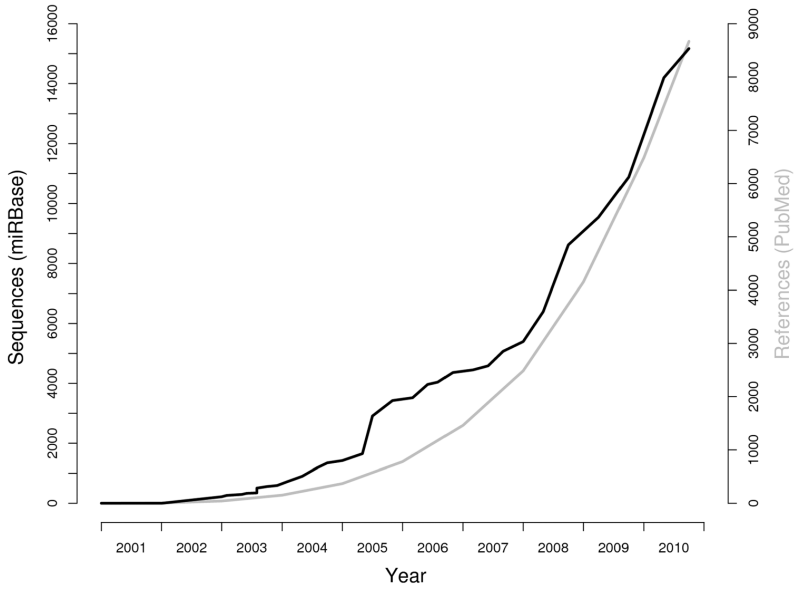


FIG. 1.4 – Nombre de microARN présent dans miRBase (en noir) et nombre de publications sur PubMed qui font référence au terme « microRNA » (en gris). Figure tirée de Kozomara et Griffiths-Jones [22].

1.2.2 Dynamique évolutive des gènes de microARN : apparition, duplication et conservation

Le nombre important de gènes de microARN par espèce ainsi que leur relative conservation ont permis d'étudier leurs caractéristiques. Nozawa *et al.* en 2010 ont comparé les microARN de 12 espèces de *Drosophila* [25]. À partir d'un catalogue de microARN homologues défini sur *D. melanogaster* et d'un arbre phylogénétique sur ces 12 espèces, si un microARN est présent chez deux espèces alors ils associent ce microARN à l'ancêtre commun de ces deux espèces. Ceci leur permet de définir un « âge » d'apparition pour chaque microARN, même si il y a un fort biais induit par le fait que le catalogue des microARN initial n'est fait qu'à partir de *D. melanogaster*. Ils ont pu observer qu'environ 50 % de l'apparition de nouveaux microARN se déroulait préférentiellement dans des régions introniques de gènes codants pour des protéines. Une des explications de l'apparition préférentielle des gènes de microARN dans des introns est que le nouveau gène de microARN est co-transcrit avec le gène codant pour une protéine et ne nécessite donc pas la présence d'un promoteur propre au gène de microARN. L'apparition de ces gènes de microARN se ferait aussi dans des régions intergéniques (∼25 %). Dans une moindre mesure (∼10 %), l'apparition de nouveaux gènes de microARN peut aussi se faire par la duplication de gènes de microARN existants. Les auteurs montrent aussi que ∼30 % des nouveaux gènes au sein des clusters de microARN sont obtenus par des duplications en tandem. En comparant le taux de mutations des microARN récents et

anciens, les auteurs ont montré que les microARN les plus anciens possédaient un taux de mutation beaucoup plus bas que les microARN nouvellement apparus et que ce taux était d'ailleurs plus faible dans la région de la graine du microARN mature. Cela suggère que les microARN ont gardé leurs fonctions au cours de l'évolution. À partir de ces observations, ils émettent l'hypothèse que lors de l'apparition de nouveaux microARN, nombre d'entre eux ne sont pas fonctionnels et disparaissent. Cependant si le nouveau gène possède une fonction alors il se fixe et est conservé. En conclusion, il semble admis que la dynamique de création et disparition des gènes de microARN puisse être très forte, avec des turn-over peut-être plus importants que pour les gènes codant des ARNm. Cependant, l'existence de microARN conservés dans l'arbre du vivant (comme let7) montre que certains gènes peuvent être très stables. Enfin, une conservation de séquence ne signifie pas une conservation de fonction : un microARN muté sur la graine peut reconnaître de nouvelles cibles [26].

1.2.3 Voie de biosynthèse des microARN

La transcription et la maturation des gènes de *microARN intergéniques* chez les insectes se fait en quatre étapes [17] qui sont illustrées Figure 1.3. Les gènes des microARN sont tout d'abord transcrits par le complexe ARN polymérase II en pri-microARN [27], brin d'ARN pouvant faire une taille de plusieurs centaines de nucléotides contenant une structure en tige-boucle d'environ 70 nucléotides et dont la tige constitue les futurs microARN matures (potentiellement fonctionnels). Ce pri-microARN est ensuite reconnu dans le noyau par un complexe protéique appelé le complexe microprocesseur qui contient deux protéines particulières : Drosha - une ARNase III - et Pasha - une protéine se fixant sur de l'ARN double brin (remplacé par DiGeorge Syndrome Critical Region 8, DGCR8 chez les mammifères et *Caenorhabditis elegans*). Ce complexe protéique va cliver les régions situées autour de la structure en tige-boucle pour former le pré-microARN : une tige-boucle où la tige fait environ 30 nucléotides pour un brin et possède en 3' deux nucléotides non appariés. Il est à noter que le pri-microARN peut contenir plus d'un gène de microARN [28] ; on parle dans ce cas de pri-microARN polycistronique et chacune des tiges-boucles du pri-microARN polycistronique (les gènes de microARN) formeront un pré-microARN unique après clivage de ces structures par le complexe microprocesseur. Une fois le pré-microARN obtenu, il est transporté du noyau vers le cytosol par l'Exportin-5, une protéine se fixant à l'ARN double brins et RanGTP dépendante [29].

Une fois dans le cytosol, le pré-microARN est pris en charge par un autre complexe protéique constitué de Dicer-1 (Dcr-1) - une ARNase III - et Loquacious. À noter que chez les insectes Dcr existe en deux copies : Dcr-1 est spécifique à la maturation des microARN et Dcr-2 spécifique à la maturation des pARNi, petits ARN interférents (en anglais : small interfering RNA, siRNA). Après que le complexe a pris en charge le pré-microARN, la boucle est clivée par Dcr-1 pour former le duplexe ARN double brins, le reste de la tige, où chaque brin possède une taille approximative de 22 nucléotides. Ce complexe contient les deux microARN matures potentiellement fonctionnels. Il est à noter que l'influence d'autres protéines a été mise en évidence dans ces différents processus de maturation [30, 31]. L'un de ces microARN matures est pris en charge par le complexe protéique RISC (RNA induced silencing complex), complexe appelé miRISC et principalement formé par des protéines de la famille Argonaute (Ago). Les

protéines de la famille Ago peuvent être divisées en deux sous catégories : Ago et Piwi où la sous-catégorie Ago contient les protéines Ago-1 et Ago-2 qui font partie du complexe miRISC. La sous-famille Piwi regroupe des protéines principalement exprimées dans les cellules germinales et qui prennent en charge d'autres petits ARN, les ARNpi (piRNA) [32]. Le chargement de l'un des deux microARN mature dans le complexe Ago-1/miRISC ou Ago-2/miRISC dépend de certaines caractéristiques du duplex ARN double brins [33]. À l'origine, ces deux microARN étaient appelés microARN mature pour le microARN fonctionnel et microARN* pour le second microARN, qui lui n'était pas considéré comme fonctionnel. Néanmoins, des preuves suggèrent que ces deux microARN matures peuvent être fonctionnels [34, 35]. Il a donc été décidé de considérer les deux brins de la tige comme des microARN matures potentiellement fonctionnels et de les nommer 5p et 3p selon leur position sur le pré-microARN, notation effective dans les dernières version de miRBase [23].

Les gènes de microARN peuvent être localisés dans des régions intergéniques mais aussi dans des gènes codant des ARNm. Dans ce manuscrit, on appellera « ARNm hôte » ou « gène hôte » des ARNm ou gènes contenant un ou plusieurs gènes de microARN. Parmi les microARN intragéniques, on peut différencier les gènes exoniques et les gènes introniques. Il a été montré que les gènes de microARN au sein d'introns des gènes codant pour des protéines étaient eux aussi exprimés [36]. L'épissage de ces microARN peut se faire de plusieurs façons. Soit l'intron du « gène hôte », relâché après épissage, forme un pri-microARN pris en charge par le complexe Drosha/Pasha pour former le pré-microARN [37]. Soit l'intron fait la taille du pré-microARN et il peut dans ce cas être directement exporté dans le cytosol sans nécessiter l'intervention d'épissage ou du complexe Drosha/Pasha. Ce type de gènes de microARN introniques sont appelés mirtrons [38]. À noter que des gènes de microARN ont aussi été détectés dans de longs ARN non codants (lARNnc) [36].

1.2.4 Mode d'action des microARN : comment ciblent-ils et régulent-ils les ARNm

Une fois le complexe miRISC obtenu, le microARN mature va servir de « guide » au complexe miRISC pour sa fixation sur l'ARNm cible à l'aide d'une complémentarité partielle entre ce microARN mature et l'ARNm. La fixation du complexe se fait préférentiellement sur le 3'UTR des ARNm même si d'autres sites de fixation ont été identifiés dans les 5'UTR ou dans les cadres ouverts de lecture [17]. La fixation préférentielle de miRISC sur le 3'UTR peut s'expliquer par le fait que si le complexe se fixe dans le 5'UTR ou le cadre ouvert de lecture, le ribosome et les autres complexes impliqués dans la traduction risquent de déplacer/détacher le miRISC de l'ARNm [16]. La fixation du miRISC peut avoir différents effets sur l'ARNm : soit inhiber sa traduction, soit induire sa dégradation, soit les deux.

Inhibition de la traduction

Il a été montré chez l'humain que l'inhibition de la traduction par miRISC pouvait se faire à l'aide de Ago-2. Un domaine protéique chez l'Ago-2 humaine a été trouvé qui permettrait sa fixation sur la coiffe m^7G (7-methyl guanosine) sur le côté 5' des ARNm et ainsi empêcher la fixation de elF4E, protéine permettant d'initier la traduction en

recrutant le ribosome [39]. Ce motif a été aussi retrouvé chez la protéine Ago-1 de *Drosophila melanogaster* ainsi que chez les Ago des chordés et *Caenorhabditis elegans* [39]. Ce domaine, non présent chez les protéines de la sous-famille Piwi, serait donc l'un des modes d'inhibition de la traduction de l'ARNm par Ago/miRISC.

Dégradation de l'ARNm

Contrairement aux plantes, très peu de microARN induisent la dégradation de leurs ARNm cibles par clivage de ces derniers chez les animaux. Chez les animaux, et donc les insectes, la dégradation de l'ARNm se fait préférentiellement par leur déstabilisation. Il a été montré chez plusieurs organismes dont *D. melanogaster* que cette déstabilisation est induite par le complexe formé de la protéine Ago-1 du miRISC et du recrutement potentiel de GW182. Ce nouveau complexe va à son tour potentiellement recruter deux complexes protéiques : CCR4/NOT et DCP1/DCP2, respectivement un complexe de déadénylation de la queue poly(A) et un complexe de décoiffage de l'ARNm. Le perte partielle de sa queue et de sa coiffe va déstabiliser l'ARNm ce qui va induire sa dégradation [40].

Inhibition et dégradation

Il a été montré que l'inhibition de la traduction et la déstabilisation de l'ARNm induite par la fixation du miRISC peuvent être conjointes et que la baisse du niveau de traduction de l'ARNm est antérieure à sa dégradation [41]. L'effet final de la fixation d'un microARN sur une cible d'ARNm est – quel que soit le mécanisme – une diminution de la traduction de l'ARNm cible. Cette effet est combinatoire puisque d'une part un microARN peut se fixer sur plusieurs cibles différentes et d'autre part un ARNm peut être cible de plusieurs microARN. Enfin, un ou plusieurs microARN matures peuvent se fixer plusieurs fois sur le même ARNm, intensifiant son effet biologique si les sites de fixation sont à une distance permettant la coopération entre ces sites (entre 15 et 35 nucléotides) [42, 16].

L'émergence d'un nouveau paradigme : la répression des microARN par les ARNm

Le paradigme classique sur les microARN est qu'ils régulent fonctionnellement un ensemble d'ARNm en diminuant leur traduction/stabilité. Depuis quelques années, il a été observé que les microARN pouvaient eux-mêmes être régulés par des ARN non codants (ARNnc) comme de longs ARN non codants (lARNnc), des pseudogènes ou encore des ARN circulaires (ARNcirc) [43, 44]. Ces observations ont mené à un nouveau paradigme : les microARN ne posséderaient qu'un ensemble très restreint de « cibles » fonctionnelles. Les autres cibles des microARN auraient pour fonction de limiter la fixation de microARN sur les véritables cibles. Ainsi, beaucoup des microARN seraient en réalité « titrés » par des ARNm ou d'autres ARNnc pour les empêcher d'être actifs sur des cibles réelles. Ces « fausses » cibles joueraient le rôle « d'éponges » à microARN [45]. Cette compétition entre microARN, ARNm, pseudogènes et lARNnc endogènes a été unifiée par une hypothèse globale appelée « compétition entre ARN endogènes »

(cARNe) ou en anglais : competing endogenous RNA (ceRNA) en 2011 par Salmena *et al.* [46].

1.2.5 Prédiction des cibles des microARN : identification des cibles et paradoxe entre le nombre élevé de prédictions et le nombre restreint d'interactions identifiées

Depuis l'identification des microARN, de leur implication dans différentes fonctions biologiques et de leurs modes d'actions sur les ARNm, un grand nombre de méthodes ont été développées afin de prédire les sites de fixation des microARN au sein des ARNm [47, 48, 49]. Cette sous-partie décrit de façon globale les critères utilisés par différentes méthodes de prédiction de sites de fixation de microARN matures sur les ARNm, et plus particulièrement leur 3'UTR. La liste ainsi que la description des méthodes ne sont pas exhaustives. De nombreux articles de synthèse décrivent et/ou comparent les différentes méthodes existantes et on pousse à se référer notamment à [47, 48, 49] pour une introduction plus approfondie.

Conservation des sites de fixation

Les sites de fixation des microARN matures sur les ARNm sont souvent conservés [16]. Cette conservation serait à la base de la conservation de la fonction des microARN puisque une perte du site de fixation entraîne une perte de l'interaction et par conséquent, une perte de fonction qui, si elle est adaptative, diminuerait la fitness de l'individu. Pour cette raison, de nombreuses études et programmes de prédiction d'interactions soit ne prédisent que les interactions où les sites de fixation sont conservés chez des espèces proches, soit permettent de le faire. Néanmoins, l'apparition de nouveaux sites de fixation n'est pas à exclure et certains de ces sites ont été montrés comme étant fonctionnels [50, 16]. De plus, lorsque que l'on étudie un caractère chez une espèce qui n'est pas partagé par d'autres, exclure ces sites de fixation non conservés qui sont donc spécifiques d'une espèce peut exclure des interactions qui seraient justement impliquées dans le caractère étudié.

Critères pour la prédiction de sites de fixation

La plupart des méthodes utilisent les critères suivants :
- La complémentarité entre le microARN mature et le 3'UTR de l'ARNm ;
- La force du duplexe formé entre le microARN mature et le site de fixation ;
- L'accessibilité du site de fixation.

Concernant la complémentarité entre le microARN mature et le 3'UTR de l'ARNm, les méthodes (par exemple TargetScan [51]) ne considèrent généralement que les couples microARN/ARNm pour lesquels il existe une complémentarité parfaite ou quasi parfaite entre la graine du microARN mature et certains types de sites de fixation (voir Figure 1.5) [16]. Même si certaines interactions existent entre des microARN matures et des ARNm où la complémentarité entre la graine et le 3'UTR n'est pas parfaite, réduire les prédictions à celles possédant ce type de complémentarité permet de réduire le taux de fausses prédictions. Pour la force du duplexe entre microARN mature et 3'UTR, l'énergie libre de la liaison, ou le changement d'énergie libre, est calculée pour définir la

stabilité de l'interaction [49]. En effet, le 3'UTR adoptant une conformation secondaire et tertiaire, le site de fixation peut être inclus dans une structure et donc ne pas être accessible pour la fixation du complexe. L'énergie nécessaire pour « ouvrir » cette partie du 3'UTR est donc importante dans l'accessibilité du site. Les critères d'accessibilité du site peuvent être très différents entre les méthodes. Certaines méthodes calculent l'énergie nécessaire pour rendre le site de fixation du 3'UTR de l'ARNm accessible au complexe miRISC comme la méthode PITA [52]. La méthode TargetScan [51] calcule le pourcentage en A/U autour du site de fixation. Les liaisons A/U étant plus faibles que les liaisons G/C, un pourcentage élevé en A/U autour du site de fixation facilitera la fixation du complexe miRISC. En plus de ces différents critères, certaines méthodes comme miRanda-mirSVR [53], DIANA-microT-CDS [54, 55], SVMicrO [56] utilisent l'apprentissage supervisé afin d'aider à la prédiction des sites de fixation. Les données d'apprentissage utilisées sont généralement des données d'expressions de microARN matures et d'ARNm. Un autre critère qui est utilisé par la majorité des méthodes est la conservation des sites de fixation chez des espèces proches afin de ne prédire que des sites fonctionnels et de réduire le nombre de faux positifs prédits.

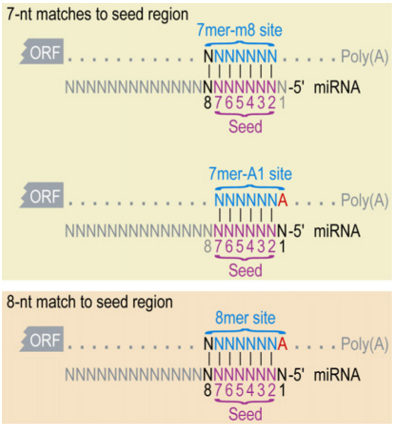


FIG. 1.5 – Les différents sites de fixation considérés comme canoniques par la méthode TargetScan. Une complémentarité parfaite des nucléotides deux à sept du microARN mature sur le 3'UTR de l'ARNm ainsi que : soit une complémentarité supplémentaire du huitième nucléotide du microARN mature (7mer-m8), soit une adénosine en aval du site de fixation sur le 3'UTR (7mer-A1), soit ces deux critères (8mer). Image adaptée de Grimson *et al.* [51].

Sur l'ensemble des méthodes développées et publiées permettant de prédire ces sites de fixation, certaines comme miRanda-mirSVR ou DIANA-microT-CDS ne sont uniquement accessibles que sous forme de base de données de prédiction chez l'humain, la souris ou encore la mouche et restent donc d'emploi limité sur des espèces non modèles. Des méthodes comme miRanda [57], PITA ou TargetScan peuvent elles être utilisées sur des données personnelles.

Un nombre élevé de fausses prédictions

Du fait de la taille réduite chez les animaux de la région d'interaction (la graine), les prédictions réalisées peuvent comprendre un grand nombre de faux positifs. Witkos

et al. en 2011 [47] ont publié une comparaison de différentes méthodes de prédiction de sites de fixation qui étaient disponibles à cette date. Leur étude comprend les méthodes miRanda, TargetScan, TargetScanS (une version antérieure à TargetScan) [58], DIANA-microT (une ancienne version de DIANA-microT-CDS) [59], PicTar (l'une des premières méthodes de prédiction) [60, 61], PITA et rna22 [62]. L'ensemble de ces méthodes a été comparé par Witkos *et al.* sur la base de trois jeux de données précédemment publiés : un premier par Sethupathy *et al.* [63] où les prédictions des méthodes sont comparées à des cibles validées expérimentalement chez les mammifères issues de TarBase [64] ; un deuxième par Beak *et al.* [65] où ils quantifient à l'aide de spectrométrie de masse la quantité de protéine en fonction de la sur-expression ou de l'inhibition de mir-223 chez la souris ; un troisième jeu par Alexiou *et al.* [66] où la quantité de protéines est évaluée en fonction de l'expression de cinq microARN. Sur la première étude, miRanda, TargetScanS et PicTar obtiennent la meilleure sensibilité avec 49 % pour miRanda et 48 % pour les deux autres. TargetScan obtient 21 % et DIANA-microT 10%. PITA et rna22 n'ont pas de sensibilité associée. La faible sensibilité de TargetScan et DIANA-microT pourrait être expliquée par le fait que ces méthodes sont plus strictes mais comme le calcul de la précision n'est pas donnée, il est difficile de conclure. Pour la seconde étude, pour toutes les méthodes exceptées TargetScanS et DIANA-microT, le changement moyen en quantité de protéine pour les cibles prédites est donné. TargetScan et PicTar obtiennent les meilleures valeurs. Pour la dernière étude, TargetScan, TargetScanS, PicTar et DIANA-microT possèdent les meilleurs précisions, entre 48 % et 51 % avec TargetScan qui possède le plus fort pourcentage. MiRanda, PITA et rna22 possèdent de plus faibles précisions, entre 24 % et 29 %. Pour la sensibilité, miRanda possède le pourcentage le plus élevé, avec 20 % alors que les autres méthodes possèdent des pourcentages autour de 10 %. Sur la dernière étude, TargetScan possède à la fois la meilleure précision et à la fois la sensibilité la plus élevée après miRanda avec un pourcentage de 12 %.

Les résultats obtenus par les méthodes de prédiction peuvent parfois être vérifiés par des tests de fixation. Par exemple, un ou plusieurs microARN matures sont transfectés dans des cellules et l'expression de cibles prédites peut être suivie au niveau de l'ARNm ou de la protéine codée. Ces approches permettent notamment de régler les paramètres des méthodes de prédiction en termes de sensibilité et de précision. Ce test biologique n'est cependant pas parfait car il ne prend pas en compte la complexité des interactions biologiques dans les cellules, notamment le fait que plusieurs ARNm peuvent être cibles d'un même microARN, ou que les microARN peuvent eux-mêmes être régulés transcriptionnellement, notamment par des facteurs de transcription [57, 67, 68].

1.2.6 Donner du sens biologique aux prédictions : l'analyse de modules d'interactions microARN/ARNm

Les microARN et leurs cibles forment un réseau d'interactions entre microARN matures et ARNm (réseau d'interaction microARN/ARNm). Suivant la question biologique à laquelle on cherche à répondre, l'espèce étudiée ou encore les données disponibles, la taille du réseau peut varier de quelques microARN, ARNm et interactions à des dizaines de microARN et des centaines ou milliers d'ARNm et d'interactions.

Afin de donner des élements de caractérisation du réseau, il est possible de réali-

ser un enrichissement fonctionnel sur les cibles de chacun des microARN matures. Il s'agit de tester statistiquement si les ARNm qui sont ciblés par un microARN matures sont impliqués préférentiellement dans des fonctions biologiques par rapport aux autres ARNm de l'espèce considérée. Néanmoins, ce type d'analyse ne prend pas en compte les groupes d'interactions impliquant plusieurs microARN matures et/ou ARNm dans les réseaux, c'est-à-dire les modules d'interaction.

Liu *et al.* [69] définissent les modules d'interaction comme un ou plusieurs sous-réseaux issus du réseau d'interactions initial dans lesquels les microARN matures, les ARNm et/ou les interactions respectent certaines contraintes. Par exemple, l'extraction de modules où l'ensemble des microARN matures et leurs cibles sont anti-corrélés ou encore des modules où tous les microARN ciblent tous les ARNm (une biclique). Les méthodes permettant l'extraction de modules utilisent principalement les interactions microARN/ARNm prédites ou validées expérimentalement ainsi que des jeux de données d'expression pour ces microARN matures et ARNm. Certaines méthodes se basent aussi sur d'autres jeux de données hétérogènes.

Nous pouvons notamment citer une méthode développée par Bryan *et al.* [70] qui utilise à la fois le réseau d'interactions, les corrélations d'expressions entre microARN matures et ARNm et l'annotation fonctionnelle des ARNm. Cette méthode intègre au sein d'une même matrice les interactions et les valeurs des corrélations. Cette matrice, où les ARNm sont représentés en ligne et les microARN en colonne, est créée de la façon suivante :

- Si la corrélation est négative et qu'il y a une interaction, alors la valeur de corrélation est gardée (une inhibition directe) ;
- Si la corrélation est positive et qu'il n'y a pas d'interaction, alors la valeur de corrélation est gardée (une activation indirecte) ;
- Pour tout les autres cas, la valeur est de zéro.

Une fois cette matrice obtenue, une recherche de biclique est faite sur cette matrice et un enrichissement fonctionnel est effectué pour chacune des bicliques. Les auteurs proposent ensuite de visualiser à l'aide du logiciel miRMAP les bicliques et les enrichissements fonctionnels (exemple Figure 1.6).

La recherche de modules au sein des réseaux d'interactions microARN/ARNm avec l'intégration d'information supplémentaire permet de caractériser ces réseaux d'un point de vue fonctionnel en isolant des microARN et des ARNm qui sont clés dans les processus étudiés.

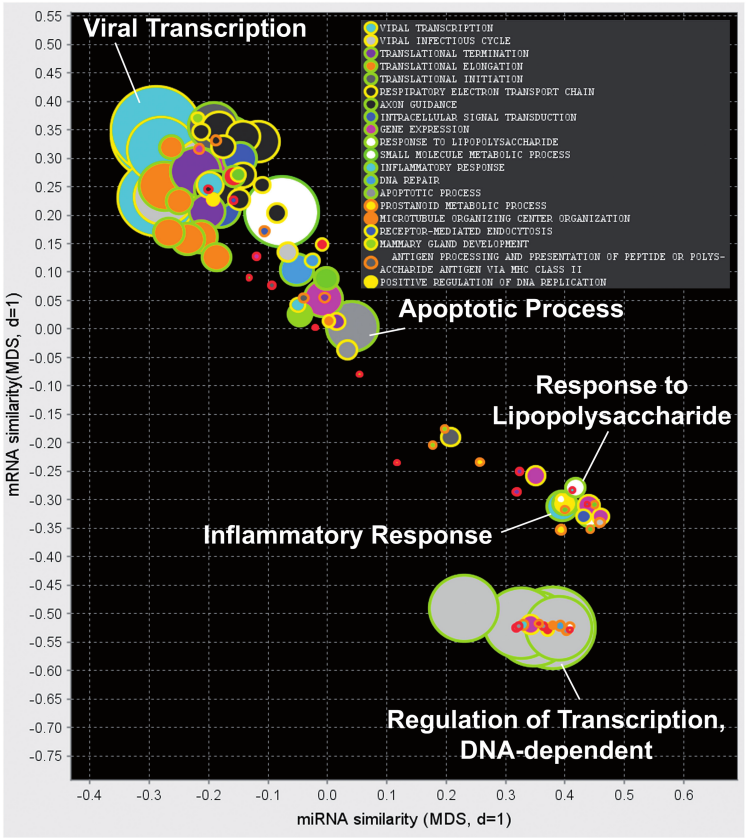


FIG. 1.6 – Exemple de visualisation de modules d'interaction par miRMAP obtenus sur un jeu de données issus de cellules immunitaires humaines. Chaque cercle représente une biclique du réseau d'interaction et les bicliques sont projetées sur un espace à deux dimensions à l'aide de multidimensional scaling (MDS). La couleur et la taille des cercles représente respectivement l'annotation Gene Ontology la plus significative ainsi que sa taille pour chacune des bicliques. Image tirée de Bryan *et al.* [70].

1.3 Les pucerons : un exemple de plasticité phénotypique

1.3.1 Description

Les pucerons appartiennent à l'ordre des Hémiptères et regroupent environ 4000 espèces différentes. Ce sont des insectes piqueurs-suceurs se nourrissant exclusivement du phloème des plantes, sève élaborée synthétisée lors de la photosynthèse et circulant dans les vaisseaux phloémiens conducteurs des tiges et nervures. Cette sève est très riche en sucre (saccharose) et acides aminés. Cependant, elle est carencée en certains acides aminés essentiels comme la phénylalanine. Au cours de l'évolution, les pucerons se sont adaptés à cette carence en développant une symbiose avec la bactérie *Buchnera aphidicola*, localisée dans des cellules particulières (les bactériocytes) dans l'abdomen des insectes, et permettant des échanges métaboliques entre ces deux partenaires de la symbiose. Les pucerons atteignent la sève phloémienne en insérant leurs pièces buccales spécialisées (les stylets) au travers des tissus végétaux environnants tels que l'épiderme ou les différents parenchymes. Les pucerons sont donc des insectes qui affaiblissent leur plante hôte en détournant des sources nutritives, et en créant des blessures sur les tissus végétaux. La sève phloémienne est également le siège de la circulation de virus de plantes qui se servent de ce fluide pour diffuser dans les différentes parties de la plante. Les pucerons peuvent ainsi absorber des virus de plantes circulant et les transmettre à des plantes saines lors de leurs différents repas. Les pucerons sont donc des vecteurs de maladies virales chez les plantes. Ainsi, les pucerons sont des ravageurs des cultures chez les plantes cultivées. Le contrôle des dégâts causés par les pucerons se fait essentiellement par l'application d'insecticides, voire par contrôle biologique en condition confinée comme des cultures sous abris (serres).

Femelle parthénogénétique vivipare

Femelle et mâle sexués

FIG. 1.7 – Photos de trois morphes du puceron du pois. À gauche une femelle parthénogénétique vivipare qui met bas à une larve. À droite l'accouplement d'une femelle et d'un mâle sexués. Crédit photo : Bernard Chaubet.

La réussite des pucerons à coloniser les cultures est due à différents traits biologiques adaptatifs, dont un mode de reproduction efficace (la parthénogenèse) et une plasticité phénotypique leur permettant de s'adapter rapidement à des changements des environnements locaux comme lors de l'alternance des saisons.

La plasticité phénotypique est la capacité d'un organisme à s'adapter à un effet

environnemental par la modification de son phénotype. Le polyphénisme est un cas particulier de la plasticité phénotypique où il y a peu de phénotypes mais qui sont très différenciés [5]. Le puceron est un modèle pour l'étude du polyphénisme. En effet, on observe chez cette espèce deux types de polyphénisme : un polyphénisme de dispersion pour lequel deux morphes adultes co-existent : ailés ou aptères, et un polyphénisme de reproduction pour lequel différents morphes sont produits soit sexués (femelles ou mâles), soit clonaux (femelles) : dans ce cas, c'est la succession des saisons qui gouverne l'apparition de ces différents morphes.

1.3.2 Le polyphénisme de reproduction du puceron

Les pucerons possèdent deux modes de reproduction, *asexué* et *sexué* avec soit des femelles parthénogénétiques vivipares (mode asexué), soit des femelles ovipares et des mâles (mode sexué). Le mode de reproduction principal des pucerons est la reproduction clonale asexuée par des femelles parthénogénétiques vivipares (à gauche Figure 1.7) : elles donnent naissance à une descendance constituée uniquement d'individus femelles qui sont génétiquement identiques entre elles et à leur mère (elles sont donc également à reproduction clonale). Ces femelles sont appelées *virginopares*. Cette viviparité est liée au développement d'embryons dans l'abdomen de la mère : les embryons les plus développés possèdent déjà leurs ovaires et les premiers embryons de la génération suivante : c'est le « télescopage » des générations. Chaque femelle clonale possède environ 80 embryons en cours de développement qui naîtront au rythme d'environ 6 par jour. Après 15 jours, ces larves seront chacune adultes et pourront à leur tour donner naissance à 80 descendants. Ce mode de reproduction permet donc un taux de natalité très élevé ce qui explique la capacité de ces insectes à coloniser leurs plantes hôtes et donc à créer des dégâts importants. Ce mode de reproduction permet d'avoir des populations où chaque individu est identique génétiquement aux autres, des populations de clones. Cependant, la viviparité chez les insectes n'est pas adaptative : privés de mécanismes de régulation de la température corporelle, les insectes supportent mal les hivers rigoureux. Les œufs hivernant sont chez les insectes souvent des formes de résistance au froid. Chez les pucerons, durant l'automne, la diminution de la durée du jour (photopériode) est perçue par les femelles comme un signal qui va modifier le développement de ses embryons : ces derniers se développeront en futurs *femelles parthénogénétiques vivipares sexupares* (en opposition avec virginopares) qui donneront à leur tour naissance à des individus sexués, mâles ou femelles (à droite Figure 1.7). Ces femelles sexuées sont ovipares et après fécondation, les œufs sont pondus. Ils sont diapausants (arrêt du développement pendant une certaine période) et n'écloront qu'au printemps suivant. Il s'agit donc d'un cas typique de plasticité phénotypique avec un signal environnemental (la photopériode) qui provoque des embryogenèses différentes aboutissant à des morphes différents et adaptés aux conditions environnementales.

Depuis l'après-guerre, des études se sont succédées pour tenter de comprendre les différentes étapes du polyphénisme de reproduction chez les pucerons [71]. La publication [71] en Annexe (6.2.2) reprend ces éléments. Ce mécanisme complexe peut être subdivisé en 3 grandes étapes : la perception du signal photopériodique, sa transduction depuis le système nerveux central jusqu'aux ovaires, et la morphogenèse des embryons. La perception de la photopériode reste mal comprise et les récepteurs non identifiés. La photopériode est perçue et intégrée en prenant en compte deux grands paramètres :

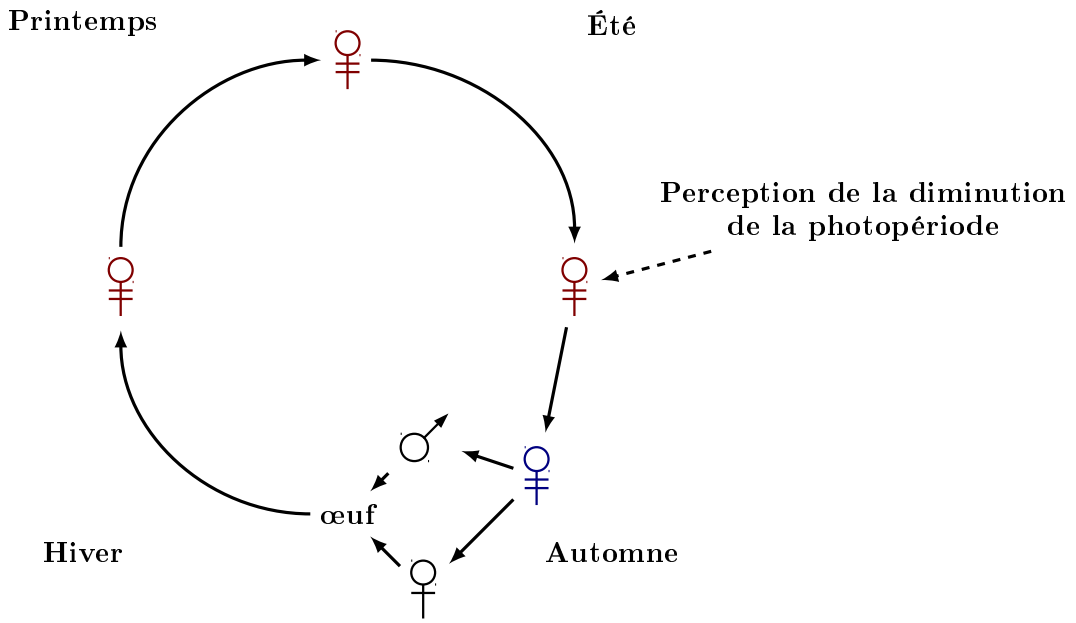


FIG. 1.8 – Schéma du cycle de reproduction du puceron du pois. En rouge, les femelles parthénogénétiques vivipares virginopares. En bleu, les femelles parthénogénétiques vivipares sexupares. En noir, les femelles et mâles sexués.

la longueur de la phase nocturne (scotophase) qui doit dépasser un seuil pour déclencher le polyphénisme, et le nombre consécutif de phases nocturnes seuillées (de l'ordre d'une dizaine de phases). Ces paramètres sont spécifiques aux différentes espèces, voire aux différentes populations. La nature du compteur moléculaire intégrant le nombre de phases nocturnes est inconnue. Plusieurs éléments de la chaîne de transduction du signal photopériodique vers les ovaires (lieu de développement des embryons) ont été identifiés comme l'importance des cellules neurosécrétrices dans le protocérébron (mais la nature des sécrétions reste inconnue) [72] ou le rôle de la mélatonine, voire de l'insuline [73, 74]. Par contre, il a été clairement montré que les hormones juvéniles étaient nécessaires à cette transduction du signal. Corbitt et Hardie en 1985 [75] ont montré que l'application sur des pucerons femelles asexuées d'hormone juvénile permettait de contrecarrer l'effet de la photopériode. Pour cela ils ont utilisé un clone d'*A. pisum*, LD 12:12, où les femelles sexupares de ce clone produisent des embryons sexués au début de la parturition (mise à bas) et après le huitième jour des embryons asexués. Ils ont appliqué de l'hormone juvénile sur l'abdomen de trois groupes différents de pucerons femelles sexupares : au début de leur stade larvaire L4, à la fin de leur stade larvaire L4 et au début de leur stade adulte. Ils ont pu observer que pour les trois groupes de pucerons, l'application de l'hormone juvénile permettait d'induire la production d'embryons asexués dans les huit premiers jours de la parturition. L'apparition de ces embryons étant plus précoce en fonction du stade lors de l'application de l'hormone : plus le puceron femelle est jeune, plus l'apparition d'embryons asexués est précoce.

Plus récemment, des travaux réalisés à l'IGEPP ont recherché les programmes génétiques chez les embryons au moment de la sélection du mode de développement sexué ou

asexué. Dans un premier temps, il a été montré que le dernier stade embryonnaire pouvant répondre aux changements environnementaux était le stade 17 (sur les 20 stades décrits, Miura *et al* [76]) : du kinoprène (molécule analogue à l'hormone juvénile) a été appliqué à la mère porteuse des embryons à son stade larvaire L4 lorsque le dernier embryon le plus mature est au stade 17. Ces mères sont élevées en photopériode courte et les embryons doivent donc se développer en futurs individus sexués. Or l'application du kinoprène contrecarre l'effet de la photopériode et ces embryons Stade 17 se développent en femelles parthénogénétique. Par contre, si le kinoprène est appliqué plus tardivement aux femelles, les embryons les plus matures (Stade 18 et au-delà) ne répondent plus au kinoprène. La fenêtre de sensibilité a donc été définie ainsi : le stade développemental 17 est le dernier pouvant modifier son programme développemental en réponse à la photopériode. Puis les stades suivants s'engagent dans le développement sexué ou asexué. Nous avons donc accès à un système expérimental permettant de suivre les modifications des programmes génétiques se mettant en place lors du polyphénisme de reproduction, permettant de comparer les embryogenèses sexuées et asexuées dans différents embryons. Le fait de travailler à condition photopériodique constante (jour court) avec ou sans kinoprène permet de plus d'assurer un développement synchrone des deux types d'embryons. Avec ce matériel, Gallot *et al.* ont, à l'aide de puces à ADN, comparé l'expression des ARNm au cours du développement d'embryons synchrones soit sexués et soit asexués [77]. Il a été montré que 33 gènes étaient différentiellement exprimés entre les trois derniers stades de développement des embryons et ces ARNm sont potentiellement impliqués dans quatre grands types de fonctions : l'ovogenèse (7 ARNm), les régulations post-transcriptionnelles (5 ARNm), les régulations épigénétiques (4 ARNm) et le cycle cellulaire (3 ARNm).

1.3.3 Le génome du puceron du pois : annotations et duplications

Ces dernières expérimentations ont été réalisées chez le puceron du pois qui est devenu depuis près de 10 ans un modèle en biologie et génomique des pucerons. En 2010, le séquençage du génome et une première annotation des gènes et transcrits d'*A. pisum* ont été publiés [15] et la base de données AphidBase a été créée afin d'avoir accès à la plupart des ressources génomiques disponibles sur *A. pisum* [78]. Depuis, une seconde annotation (2.1) des gènes a été effectuée (disponible sur Aphidbase). Le génome du puceron du pois, constitué de quatre chromosomes, fait environ 540 mega base (Mb) et le séquençage a permis d'assembler un peu plus de 20.000 « scaffolds ». Sur l'ensemble de ce génome, 36.973 gènes et 36.990 transcrits ont été identifiés. Il a été observé que parmi l'ensemble des gènes prédits, un grand nombre résultait de duplications : environ 2.000 familles de gènes dupliqués sont décrites [15] (où une famille de gènes est définie comme des gènes similaires et donc potentiellement de fonction biochimique proche). Parmi les fonctions des familles dupliquées nous pouvons citer la modification de la chromatine, le transport des sucres ou encore la synthèse des microARN (discuté plus bas). Par contre, certains gènes qui sont généralement conservés durant l'évolution sont ici manquants, comme des gènes impliqués dans la réponse immunitaire de type IMD ou certaines voies métaboliques.

Sept protéines sont impliquées dans la synthèse et la maturation des microARN : *Drosha* et *Pasha* pour le passage de pri-microARN vers le pré-microARN, l'*Exportine-5* pour le transport du pré-microARN du noyau vers le cytoplasme, *Dicer-1* et *Loqua-*

cious pour la maturation du pré-microARN en complexe ARN double brin et finalement *Argonaute-1* qui fait partie du complexe RISC impliqué dans la répression induite par les microARN matures. Sur les sept gènes codants pour ces protéines, quatre d'entre eux sont dupliqués sur le génome : Pasha en quatre exemplaires et Dicer-1, Loquacious et Argonaute-1 en deux exemplaires [79, 15]. Ces duplications ont été observées chez d'autres espèces de puceron mais ne semblent pas être présentes chez d'autres metazoa en dehors des pucerons [80]. Dans chaque cas, une des copies n'a pas évolué et reste très similaire aux orthologues des autres insectes, et la seconde copie a fortement évolué, avec des séquences éloignées des copies originales, et parfois des traces de sélection positive, évoquant l'acquisition potentielle d'une nouvelle fonction. De plus, l'expression comparative entre les différents morphes de ces gènes montre que les copies originales s'expriment de façon ubiquiste chez tous les morphes, alors que les copies divergentes acquièrent des spécificités d'expression : par exemple, Dicer-1b (copie divergente de Dicer) a un profil d'expression très marqué chez les femelles parthénogénétiques sexupares. Ces observations indiquent une potentielle spécialisation de la machinerie des microARN, et permettent de poser l'hypothèse que des régulations génétiques par des microARN pourraient intervenir lors du polyphénisme de reproduction.

1.3.4 Les différents modes de reproduction et les microARN

Deux études précédentes ont étudié l'expression des microARN chez différents morphes du mode de reproduction chez *A. pisum*. En 2010, Legeai *et al.* ont publié un premier catalogue de microARN chez le puceron du pois [81]. Ce catalogue a été obtenu à l'aide de séquençage haut débit de petits ARN chez trois types de morphes : des femelles parthénogénétiques virginopares, des femelles parthénogénétiques sexupares et des femelles ovipares. En plus de l'identification de ce catalogue, l'expression dans les différents morphes a été comparée. Cette étude a permis de mettre en évidence 17 microARN matures différentiellement exprimés entre les morphes avec notamment *api-let-7* et *api-mir-100* qui sont sur-exprimés chez le morphe ovipare comparé aux morphes virginopares et sexupares et qui sont impliqués dans la métamorphose et la réponse à l'ecdysone, une hormone du développement chez les insectes. Le microARN *api-mir-34* impliqué lui aussi dans la réponse à l'ecdysone et à l'hormone juvénile chez *Drosophila melanogaster* [82] est sur-exprimé chez le morphe sexupare comparé au morphe virginopare. En 2011, Aurore Gallot a dans sa thèse étudié l'expression des microARN lors du développement d'embryons synchrones sexués et asexués [83] sur le même modèle expérimental que celui des ARNm. Dans cette thèse, Gallot a montré que trois microARN qui avaient déjà été observés comme différentiellement exprimés par Legeai *et al.*, *api-mir-2a*, *api-mir-7* et *api-mir-275*, sont aussi différentiellement exprimés dans son étude, au niveau des embryons. L'ensemble de ces résultats sur l'expression de microARN matures chez différents morphes du puceron du pois ou chez les embryons met en valeur un rôle potentiel des microARN dans la détermination du mode de reproduction chez *Acyrthosiphon pisum*.

1.4 L'analyse de concepts formels

L'analyse de concepts formels (ACF) est un formalisme introduit par Ganter et Wille [84] qui permet de former des concepts en regroupant des objets en fonction de leurs attributs. On part de deux ensembles, un ensemble d'objets et un ensemble d'attributs, et d'une relation binaire entre ces objets et ces attributs (attributs associés à chaque objet). Chaque concept sera défini par son extension, l'ensemble d'objets sur lequel le concept s'applique, et l'ensemble d'attributs qui caractérise le concept, son intension. En plus de la formalisation des concepts, l'ACF permet d'ordonner partiellement les concepts. Nous utilisons l'ACF pour l'analyse du réseau d'interactions.

1.4.1 L'analyse de concepts formels : notation et définitions

Un *contexte formel* est un triplet $\mathbb{K} = (G, M, I)$ où G est un ensemble d'objets, M un ensemble d'attributs et $I \subseteq G \times M$ est une relation binaire entre les objets et les attributs. L'opérateur $(.)'$ est défini sur $\mathbb{K}$ pour $A \subseteq G$ et $B \subseteq M$ comme :

$$A' = \{m \in M | \forall g \in A : gIm\}; \qquad B' = \{g \in G | \forall m \in B : gIm\}.$$

A' est donc l'ensemble d'attributs communs à tous les objets de A et B' l'ensemble d'objets communs à tous les attributs de B.

Un *concept formel* est une paire (A, B) définie sur $\mathbb{K}$ avec $A \subseteq G$ et $B \subseteq M$ où $A = B'$ et $B = A'$. A est appelé l'*extension* du concept et B l'*intension* du concept. Dans la suite de la thèse, le mot concept sera utilisé pour faire référence aux concepts formels.

Les concepts peuvent être ordonnés en se basant sur l'inclusion de leurs ensembles : pour deux concepts (A, B) et (C, D), si $A \subset C$ (réciproquement $C \subset A$) alors on note $(A, B) < (C, D)$ (réciproquement $(C, D) > (A, B)$). De plus, $A \subset C$ implique $D \subset B$. Si $(A, B) < (C, D)$ et qu'il n'existe aucun concept formel (E, F) tel que $(A, B) < (E, F) < (C, D)$ alors on note que $(A, B) \prec (C, D)$ (réciproquement $(C, D) \succ (A, B)$).

La relation $<$ génère un treillis de concepts noté $\underline{\mathfrak{B}}(\mathbb{K})$ sur le contexte $\mathbb{K}$. L'ordre $\prec$ représente les arêtes du graphe couvrant $\underline{\mathfrak{B}}(\mathbb{K})$. Ces relations permettent de définir l'*infimum* de deux concepts, c'est-à-dire la plus grande borne inférieure de deux concepts et le *supremum* de deux concepts, la plus petite borne supérieure de deux concepts. Dans la suite de la thèse, lorsqu'il sera fait référence au treillis de concepts il sera fait référence au graphe couvrant $\underline{\mathfrak{B}}(\mathbb{K})$ généré par la relation $\prec$. Sur le treillis, on peut définir un concept maximum appelé top ($\top$). L'extension de ce concept est l'ensemble G des objets du contexte formel et son intension est l'ensemble éventuellement vide des attributs possédés par l'ensemble des objets du contexte formel. Réciproquement on peut définir un concept minimum, le concept bottom ($\bot$). L'intension de ce concept est l'ensemble M des attributs du contexte formel et son extension est l'ensemble éventuellement vide des objets qui possèdent l'ensemble des attributs du contexte formel.

Le Tableau 1.1 présente un exemple de contexte formel et le treillis associé est présenté Figure 1.9. Ce contexte contient quatre concepts : C_1, C_2, et les concepts $\top$ et $\bot$.

Les concepts formels peuvent être vus comme des rectangles maximaux de 1 dans la relation binaire modulo des permutations de lignes et/ou de colonnes. Des rectangles

	a_1	a_2	a_3	a_4
o_1	1	1		
o_2	1	1		
o_3	1	1		
o_4			1	1
o_5			1	1

Tableau 1.1 – Contexte formel $\mathbb{K}_{\text{ex}} = (G_{\text{ex}}, M_{\text{ex}}, I_{\text{ex}})$ avec $G_{\text{ex}} = \{o_1, .., o_5\}$ l'ensemble des objets et $M_{\text{ex}} = \{a_1, .., a_4\}$ l'ensemble des attributs. Les 1 représentent la relation binaire I_{ex}.

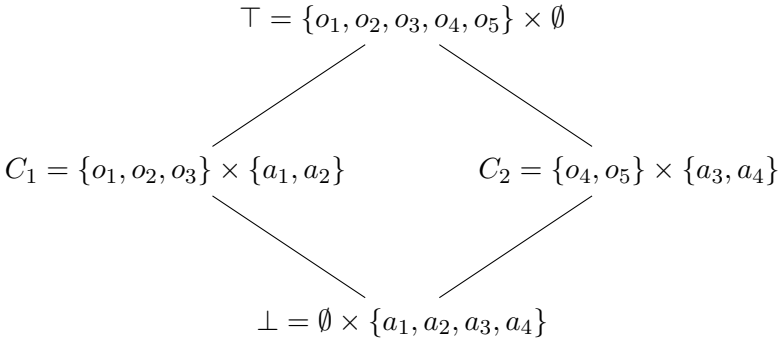


FIG. 1.9 – Treillis de concepts $\underline{\mathfrak{B}}(\mathbb{K}_{\text{ex}})$ du contexte formel $\mathbb{K}_{\text{ex}}$. Les concepts sont présentés de la façon suivante : nom du concept $= A \times B$ avec A l'ensemble des objets du concept et B l'ensemble des attributs du concept.

sont dit maximaux si l'on ne peut rajouter aucune colonne avec des 1 sur l'ensemble des lignes et aucune ligne avec des 1 sur l'ensemble des colonnes. Sur l'exemple Tableau 1.1 on peut voir que les deux concepts formels C_1 et C_2 définis Figure 1.9 sont en effet des rectangles maximaux.

1.4.2 Le graphe d'interactions entre microARN et ARNm : un graphe biparti modélisable comme un contexte formel

Un graphe $G = (N, E)$ est un ensemble de nœuds N reliés par un ensemble E d'arêtes. on peut le représenter par sa matrice d'adjacence. Pour un graphe $G = (N, E)$ simple, c'est-à-dire tel qu'il existe au maximum une seule arête entre deux nœuds, la matrice d'adjacence T du graphe est une matrice booléenne de dimension $|N| \times |N|$ où la valeur de la case t_{ij} de la matrice d'adjacence T vaut :

$$t_{ij} = \begin{cases} 1 & \text{si il y a une arête entre le nœud } i \text{ et le nœud } j \text{ dans } E, \\ 0 & \text{sinon.} \end{cases}$$

Pour chaque nœud du graphe, le nombre d'arêtes qui sont reliées à lui est appelé le degré du nœud.

Les graphes bipartis sont un cas particulier de graphe où il existe une partition de l'ensemble des nœuds en deux sous-ensembles distincts U et V tel que chaque arête de

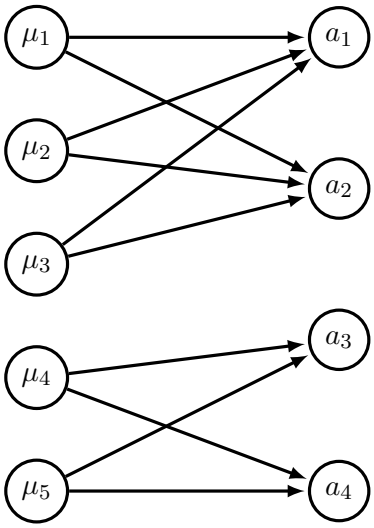
E relie un nœud de U à un nœud de V. On peut noter un graphe biparti de la façon suivante : $G = (U, V, E)$ afin de distinguer les deux ensembles de nœuds. Par exemple, un réseau d'interactions entre des microARN matures et des ARNm est un graphe biparti noté $R = (\mu, ARN, I)$ où les nœuds de μ sont l'ensemble des microARN matures, les nœuds de ARN l'ensemble des ARNm et les arêtes I l'ensemble des interactions entre les microARN matures issus de μ et les ARNm issus de ARN. En effet, les seules interactions qui existent dans ce réseau sont des interactions entre les microARN matures et les ARNm, il n'existe aucune arête entre les microARN matures et aucune arête entre les ARNm. Dans la suite de la thèse l'ensemble des microARN matures sera représenté par les lignes de la matrice d'adjacence et l'ensemble des ARNm sera représenté par les colonnes de la matrice d'adjacence.

La Figure 1.10 montre un graphe biparti $R_{\text{ex}} = (\mu_{\text{ex}}, ARN_{\text{ex}}, I_{\text{ex}})$ et sa matrice d'adjacence T_{ex} correspondante avec un ensemble d'interactions I_{ex} entre un ensemble μ_{ex} de microARN matures et un ensemble ARN_{ex} d'ARNm. On peut voir que la matrice d'adjacence T_{ex} du graphe biparti R_{ex} présentée Figure 1.10 est identique au contexte formel $\mathbb{K}_{\text{ex}}$ présenté Tableau 1.1. La matrice d'adjacence T_{ex} peut donc être modélisée comme un contexte formel où l'ensemble des objets correspond à l'un des deux ensembles de nœuds (ici l'ensemble μ_{ex} des microARN matures), l'ensemble des attributs correspond à l'autre ensemble de nœuds (ici l'ensemble ARN_{ex} des ARNm) et la relation binaire entre objets et attributs correspond à l'ensemble des arêtes du graphe (ici l'ensemble I_{ex} des interactions en microARN matures et ARNm).

En théorie des graphes, une clique est un sous-ensemble de nœuds qui sont tous reliés les uns aux autres par des arêtes. On dit d'une clique qu'elle est (de taille) maximale si l'on ne peut ajouter aucun nœud à la clique pour former une clique plus grande. Les bicliques sont des extensions de la notion de clique adaptées aux graphes bipartis : elles sont formées de deux sous ensembles de nœuds reliés entre eux. La biclique est (de taille) maximale si on ne peut ajouter aucun nœud sans perdre cette propriété. Sur les matrices d'adjacences de graphes bipartis, une biclique est représentée par un sous-ensemble de lignes et un sous-ensemble de colonnes qui forment un rectangle de 1. Dans le cas des bicliques maximales, ces rectangles sont eux aussi maximaux. De la même façon que les matrices d'adjacences de graphes bipartis sont des contextes formels, les bicliques de tailles maximales sont en faites des concepts formels.

1.4.3 L'analyse de concepts formels en bio-informatique

L'ACF est utilisée pour de nombreuses applications que ce soit en fouille de données, fouille de texte, web sémantique ou encore en biologie. Concernant la biologie, des méthodes ont été développées principalement sur des puces à ADN [85, 86, 87, 88]. D'autres méthodes utilisent aussi l'ACF pour compléter des réseaux de régulation de gènes à l'aide de cinétiques d'expression [89] ou encore pour explorer des dépendances temporelles sur des réseaux biologiques booléens [90]. Même si l'ACF peut s'apparenter à du biclustering, elle a l'avantage de générer l'ensemble des regroupements possibles basés sur le relation binaire entre objets et attributs et non pas comme la plupart des méthodes de biclustering un seul partitionnement des données, cherchant à optimiser une certaine fonction objectif. Ceci permet, par exemple, d'avoir des gènes/ARNm qui appartiennent à plusieurs groupes, les concepts, dans le cas d'étude de puces à ADN, ce qui n'est pas obtenu par du biclustering classique où chaque élément appartient à



	a_1	a_2	a_3	a_4
μ_1	1	1		
μ_2	1	1		
μ_3	1	1		
μ_4			1	1
μ_5			1	1

FIG. 1.10 – Exemple, à gauche, d'un graphe biparti $R_{\text{ex}} = (\mu_{\text{ex}}, ARN_{\text{ex}}, I_{\text{ex}})$ d'interactions entre des microARN matures ($\mu_{\text{ex}} = \{\mu_{1..5}\}$) et des ARNm ($ARN_{\text{ex}} = \{a_{1..4}\}$) et, à droite, la matrice d'adjacence T_{ex} associée.

un et un seul groupe. Ceci permet d'identifier des éléments impliqués dans différents processus, ce qui est la cas en biologie. En plus de ce chevauchement de concepts, leur hiérarchisation par le treillis permet d'observer des ensembles qui seraient « génériques », comme par exemple des facteurs de transcription impliqués dans un grand nombre de processus, et des ensembles qui seraient spécifiques d'une condition, des facteurs de transcription impliqués dans très peu de processus biologiques.

1.5 L'Answer Set Programming

Dans la thèse, de nombreux codes ont été écrits en langage Answer Set Programming (ASP) [91, 92] et utilisés à l'aide de la suite logicielle Potassco [93].

L'ASP est une forme de programmation logique par contraintes issu du domaine de la représentation des connaissances et qui permet de résoudre un problème non pas en décrivant la façon de le résoudre, l'algorithmique, mais en décrivant la spécification du problème à l'aide de déclarations logiques. Un moteur de résolution (solver) va ensuite trouver l'ensemble des modèles qui satisfont le problème posé sans que l'on ait à se préoccuper de la manière dont le solver va trouver ces modèles. En plus de l'énumération de l'ensemble des modèles, certains solvers comme *clasp* [94] permettent d'obtenir l'intersection de l'ensemble des modèles, leur union ou encore de trouver des modèles maximaux ou minimaux suivant des critères définis par l'utilisateur.

L'idée est de formuler la connaissance par des faits logiques puis de raisonner sur ces faits logiques en définissant des ensembles de règles et de contraintes.

Les règles sont de la forme : $t :- c_1,\ldots,c_n.$ où t est l'atome de la tête de la règle, $c_1, \ldots, c_n$ les atomes du corps. Le sens d'une règle (sémantique) est que l'atome de tête est nécessairement vrai si tous ceux du corps le sont.

Les contraintes sont de la forme : $:- c_1,\ldots, c_n.$ avec la sémantique qu'au moins un des c_i doit être faux. Sinon le modèle est faux. Un exemple d'un code ASP est donné (Figure 1.11) qui renvoie la ou les plus grandes bicliques maximales du graphe biparti présenté Figure 1.10.

Les atomes des lignes 2 à 6 représentent les faits, ici l'ensemble des interactions du graphe. Les lignes 11 et 13 définissent respectivement les microARN et les ARNm du graphe à partir des interactions : si un élément est dans une interaction alors c'est un nœud du graphe. La ligne 16 permet de choisir les microARN du graphe appartenant à la biclique où cMicro(X) est vrai si X a été choisi. L'expression :micro(X) permet de définir le domaine sur lequel cMicro(X) est défini. La ligne 17 définit la même chose mais pour les ARNm. La ligne 23 définit les microARN impossible, c'est-à-dire les microARN qui ne peuvent pas faire partie de la biclique car ils ne possèdent pas d'interaction pour au moins un ARNm choisi. La ligne 24 définit la même chose pour les ARNm. La ligne 28 est la contrainte sur les interactions au sein de la biclique : le modèle est rejeté si on a choisi le microARN X et l'ARNm Y et qu'il n'y a pas d'interaction entre X et Y. La ligne 34 est la contrainte sur la maximalité de la biclique : le modèle est faux si un microARN possède une interaction avec tous les ARNm de la biclique mais n'en fait pas partie. La ligne 35 est la même contrainte sur les ARNm. Finalement, la ligne 39 permet d'obtenir la ou les plus grandes bicliques maximales : les modèles qui maximisent le nombre de microARN et d'ARNm choisis. À noter que pour avoir l'ensemble des bicliques maximales, et non plus uniquement celles de plus grandes tailles, il suffit d'enlever la ligne 39.

```
1  % Définition des faits : les interactions du réseau
2  inter("mu1","a1"). inter("mu1","a2").
3  inter("mu2","a1"). inter("mu2","a2").
4  inter("mu3","a1"). inter("mu3","a2").
5  inter("mu4","a3"). inter("mu4","a4").
6  inter("mu5","a3"). inter("mu5","a4").
7
8  % X est un microARN ou un ARNm du réseau
9  % si il est dans une interaction
10 % Pour les microARN
11 micro(X) :- inter(X,_).
12 % Pour les ARNm
13 arn(X)   :- inter(_,X).
14
15 % On choisit au moins 1 élément dans chacun des ensembles
16 1{cMicro(X):micro(X)}.
17 1{cArn(X):arn(X)}.
18
19 % Le microARN X est considéré comme ne pouvant pas
20 % faire partie de la biclique (imp) si il ne possèdent
21 % pas une interaction avec tous les ARNm de la biclique
22 % Pareil pour les ARNm
23 impMicro(X) :-  micro(X), cArn(Y), not inter(X,Y).
24 impArn(Y)   :- cMicro(X),  arn(Y), not inter(X,Y).
25
26 % Pour avoir une biclique, on ne peut pas avoir choisi
27 % deux nœud X et Y et qu'ils ne soient pas en interaction
28 :- cMicro(X), cArn(Y), not inter(X,Y).
29
30 % Pour que la biclique soit maximale, tout les
31 % microARN qui ne sont pas impossible doivent
32 % faire partie de la biclique
33 % Pareil pour les ARNm
34 :- micro(X), not impMicro(X), not cMicro(X).
35 :-   arn(Y), not   impArn(Y), not   cArn(Y).
36
37 % La ou les plus grandes bicliques maximales sont les
38 % bicliques avec le nombre maximum de microARN et ARNm
39 #maximize {cMicro(X), cArn(Y)}.
```

FIG. 1.11 – Exemple d'un code ASP permettant de trouver la biclique maximale sur le réseau Figure 1.10. Les commentaires sont précédés par le caractère %.

1.6 En résumé

Il a été mis en évidence dans cette introduction l'influence potentielle des microARN sur l'expression des ARNm et leur probable implication dans la plasticité du mode de reproduction chez *Acyrthosiphon pisum*. La thèse cherche à discriminer au niveau génomique entre le développement d'embryons vers un mode de reproduction sexué et le développement vers un mode asexué. Cette discrimination passe par la création du réseau de régulation post-transcriptionnelle des microARN et des ARNm qui possèdent des cinétiques d'expression différentes entre ces deux embryogenèses et l'analyse des modules d'interactions de ce réseau par l'utilisation de l'ACF.

Pour ce faire, une stratégie en cinq étapes a été mise en place :

1. Création d'un catalogue à jour des microARN chez *A. pisum* à l'aide de données de séquençage haut-débit (Chapitre 2) ;
2. Création d'un réseau d'interactions entre les microARN prédits précédemment et les ARNm du puceron du pois (Chapitre 2) ;
3. Extraction et réduction du réseau aux microARN et ARNm qui possèdent des cinétiques différentes entre les deux embryogenèses à partir des données d'expression tirées du séquençage haut-débit (Chapitre 3) ;
4. L'utilisation de l'ACF dans le cadre de l'étude de réseaux bipartis d'interactions microARN/ARNm appliquée à *A. pisum* (Chapitre 4) ;
5. Analyse du réseau d'interactions réduit aux éléments d'intérêt par l'analyse de concepts formels (Chapitre 5).

L'étude du réseau nous donnera des pistes pour la validation des interactions microARN/ARNm prédites et pour des expériences pouvant venir enrichir le réseau. De plus, toutes ces études nous permettront de pouvoir formuler des hypothèses quant à la différenciation du mode de reproduction chez le puceron du pois.

Chapitre 2

Catalogues des ARNm et des microARN du puceron du pois *Acyrthosiphon pisum*

À l'occasion de la publication du génome du puceron du pois *Acyrthosiphon pisum* en 2010, une première version des catalogues des gènes, des transcrits d'ARNm et de microARN a été publié [15, 81]. Depuis, une nouvelle version du génome du puceron du pois est disponible (v2.1 *Acyrthosiphon pisum* LSR1) et a nécessité que nous mettions à jour la description à la fois des ARNm et des microARN. Ce chapitre présente brièvement le catalogue des ARNm et détaille le catalogue des gènes codant les microARN. De plus, cette partie présente pour la première fois le catalogue des interactions prédites entre microARN matures et transcrits d'ARNm.

2.1 Matériels et méthodes : extraction, séquençage et analyse des longs et petits ARN

2.1.1 Description des ARNm du puceron du pois

L'annotation du génome du puceron du pois a conduit à un Catalogue Officiel des Gènes (Official Gene Set) dont nous reprenons la nomenclature et à partir duquel nous basons le catalogue des ARNm [15]. La première version du génome du puceron du pois a été assemblée, annotée et décrite une première fois en 2010 par The International Aphid Genomics Consortium [15]. Depuis, l'assemblage et l'ensemble des annotations ont été améliorés afin de fournir une seconde version de ce génome (v2[1] [78]). Ce travail utilise cette annotation.

En résumé, l'annotation de la majorité des gènes codant pour des protéines a été élaborée en utilisant le consensus de différents modèles de prédiction de gènes. Ces modèles de prédiction ce sont basés sur des gènes qui étaient déjà totalement séquencés ou des gènes partiellement séquencés pour lesquels il y avait des séquences exprimées (expressed sequence tag, EST) chez le puceron du pois. Du séquençage haut débit de longs ARN (RNA-Seq) a aussi été utilisé provenant de banques différentes.

Ces modèles ont prédit un ensemble de 36.990 transcrits et 36.973 gènes pour l'ensemble du génome d'une taille de 541,69 Mb. L'étude de l'ensemble des gènes codant pour des protéines a mis en évidence l'expansion d'un grand nombre de familles de gènes et notamment des gènes impliqués dans la synthèse et l'action des microARN. Le nombre de familles de gènes dupliqués s'élevait à 2.459 pour la première version de l'annotation du génome [15, 95].

Annotations fonctionnelles des ARNm

L'annotation fonctionnelle par les termes de la Gene Ontology (GO) des transcrits d'ARNm a été réalisée dans l'équipe Écologie et Génétique des Insectes à l'IGEPP par Fabrice Legeai et Anthony Bretaudeau à l'aide du logiciel Blast2GO qui permet d'annoter par des termes GO un ensemble de gènes ou de transcrits en repérant par homologie des séquences et des domaines protéiques connus. Pour la recherche de séquences similaires par BLAST, seules les 40 meilleurs séquences avec une E-value inférieure ou égale au seuil de 10^{-8} ont été gardées. Une prédiction de domaines protéiques a été obtenue avec InterProScan afin d'enrichir l'annotation fonctionnelle obtenue par Blast2GO. Le Tableau 2.1 présente le nombre de transcrits d'ARNm du puceron du pois possédant une ou plusieurs séquences homologues, le nombre de transcrits avec un domaine protéique et le nombre de transcrits pour lesquels une annotation fonctionnelle GO a pu être obtenue par Blast2GO.

Sur l'ensemble des 36.990 transcrits, 10.062 ont pu être annotés fonctionnellement par 41.031 annotation GO, soit à l'aide de séquences homologues (31.151 protéines avec des séquences homologues obtenues par BLASTP), soit à l'aide de l'identification de domaines protéiques (28.945 protéines avec des domaines protéiques annotés fonctionnellement à l'aide InterProScan).

[1] `http://www.aphidbase.com/`

Type	Nombre	Pourcentage
Total	36.990	100 %
Avec un homologue par BLASTP	31.151	84 %
Avec un domaine protéique par InterProScan	28.945	78 %
Avec une annotation fonctionnelle GO	10.062	27 %

Tableau 2.1 – Annotation des gènes codant des ARNm dans le génome d'*A. pisum*. Les nombres de transcrits d'ARNm avec une séquence homologue, une annotation InterProScan et une annotation fonctionnelle issues de la Gene Ontology sont indiqués.

2.1.2 Élevage des pucerons, extraction et séquençage des longs ARN et des petits ARN

Les protocoles décrits ci-dessous ont été mis au point et suivis par Gaël Le Trionnaire, Sylvie Tanguy, Sylvie Hudaverdian et Nathalie Leterme. Ils sont adaptés de ceux décrits par Aurore Gallot [77].

Élevage des pucerons

L'objectif est de produire de façon synchrone des embryons qui, adultes, seront soit des femelles parthénogénétiques, soit des femelles sexuées. La synchronie est nécessaire car les vitesses de développement de ces différents types d'embryons peuvent varier sous les différentes photopériodes courtes ou longues. Gallot et al. [77] ont adapté pour cela un protocole qui permet sous une même photopériode de produire soit des embryons sexués, soit des embryons asexués de façon synchrone. Cette synchronie est obtenue par l'application de kinoprène, molécule analogue de l'hormone juvénile et qui permet de renverser l'effet d'une photopériode courte [96]. Gallot et al. [77, 83] ont déterminé le moment d'application du kinoprène pour dévier le développement sexué attendu (photopériode courte) en développement asexué.

Le protocole est décrit sur la Figure 2.1. À cause de la viviparité, ce processus se déroule sur plusieurs générations. Des larves au stade L1 sont isolées et placées en photopériode longue (16h) jusqu'à leur mue au stade L3 : c'est la génération G0. Les L3 du même âge sont alors placées sur des plantes en photopériode courte (12h) pour enclencher le développement d'embryons sexués. Lorsque les adultes apparaissent, elles commencent à produire leurs descendants, des larves de stade L1 de la génération G1. Quelques larves de ces générations sont isolées, au même âge et placées sur de nouvelles plantes, toujours en photopériode courte.

Lorsque ces larves arrivent au stade L4 + 24h (24h après leur mue L3 vers L4), elles sont séparées en deux lots : le premier lot se voit appliquer de l'acétone (50 nL) sur leur abdomen. Le second lot se voit appliquer dans les mêmes conditions 50 nL de kinoprène (400 ng, Sigma Aldrich) dilué dans l'acétone. Les deux lots sont laissés en photopériode courte. Lorsque ces L4 deviennent adultes, elles donnent naissance à la génération 2 (G2) qui sera soit sexuée (pas d'application de kinoprène) soit asexuée (application de kinoprène). Mais les échantillons sont récoltés avant que cette G2 soit mise bas. Les femelles de la G1 sont disséquées après application de l'acétone ou du kinoprène, afin d'extraire et d'isoler les embryons (futurs G2). Quatre stades d'embryons sont disséqués : le stade 17 (L4 + 24h, au moment de l'application de l'acétone avec ou sans kinoprène),

le stade 18 (L4 + 48h), le stade 19 (Adulte + 24h) et le stade 20 (adulte + 48h). Les dissections aux stades 18, 19 et 20 sont réalisées pour les lots sans et avec kinoprène. Pour chaque dissection, seuls les embryons des stades indiqués sont disséqués, en se référant aux morphologies décrites dans Miura et al. [76]. Les différents lots d'embryons (7 au total) sont congelés immédiatement dans l'azote liquide et conservés à -80 ° C.

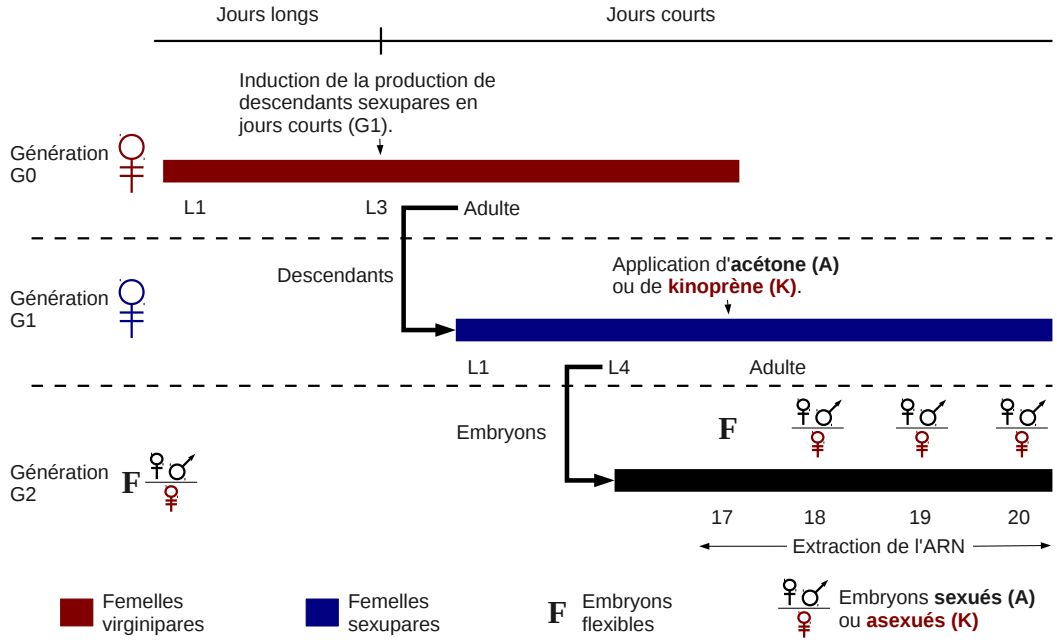


FIG. 2.1 – Schéma expérimental de la production d'embryons femelles à l'avenir sexué ou asexué à partir de femelles parthénogénétiques. La durée du jour, long ou court, est indiquée en haut ; la génération G0 est représentée en haut en rouge ; la génération G1 est représentée au milieu en bleu ; la génération G2 est représentée en bas en noir. Le morphe de reproduction et le type de descendance sont indiqués à côté des générations G0 et G1. Pour la génération G2, l'état flexible et les deux modes de reproductions sont indiqués. Les points d'extraction des ARN sont indiqués pour les stades embryonnaires 17, 18, 19 et 20 de la génération G2. Il est à noter que le schéma n'est pas à l'échelle temporelle.

Extraction et séquençage des ARN

L'ensemble du protocole décrit précédemment a été effectué séparément pour l'extraction des longs et des petits ARN et à différentes reprises afin d'obtenir des réplicats biologiques :

- 1 réplicat biologique pour l'extraction des longs ARN au stade 17 ;
- 3 réplicats biologiques pour chacun des points d'extraction des longs ARN aux stades 18, 19 et 20 et pour chacun des deux groupes : futurs asexués et futurs sexués ;
- 3 réplicats biologiques pour chacun des points d'extraction des petits ARN au stade 17 ainsi qu'aux stades 18, 19 et 20 pour chacun des deux groupes : futurs

asexués et futurs sexués.

L'extraction à partir des échantillons congelés et le séquençage des ARN ont été effectués de la façon suivante :

- Longs ARN, stade 17 : Les ARN ont été séquencés sur une machine Illumina Hiseq 2000 avec des lectures paired end. De plus, le séquençage a été effectué sur 3 pistes différentes, donnant lieu à 3 réplicats de séquençage pour le même échantillon ;
- Longs ARN, stades 18, 19 et 20 : à l'aide du kit RNeasy (Qiagen) en suivant les instructions du fabricant. Ces ARN (2 μg) ont été séquencés sur une machine Illumina Hiseq 2000 avec des lectures paired end et orientées ;
- Petits ARN stades 17, 18, 19 et 20 : Les ARN totaux enrichis en petits ARN ont été extraits à l'aide du kit miRVANA (Life Technologies). Le protocole a été réalisé en suivant les instructions du fabricant. Les petits ARN ont été sélectionné dans 5 μg d'ARN puis séquencés sur une machine Illumina HiSeq 2000 avec des lectures orientées de 50 nucléotides. De plus, chaque réplicat biologique a été séquencé sur 3 pistes différentes, donnant lieu à 3 réplicats de séquençage pour chacun des réplicats biologiques.

En résumé, le nombre total de banques de séquences obtenu est le suivant :

- 3 banques pour le stade 17 pour les longs ARN ;
- 18 banques pour les stades 18, 19 et 20 pour les longs ARN dont 9 banques pour le groupe des futurs asexués et les 9 autres pour le groupe des futurs sexués ;
- 63 banques pour les stades 17, 18, 19 et 20 pour les petits ARN dont 9 banques pour le stade 17 où les embryons sont flexibles, 27 banques pour le groupe des futurs asexués et les 27 autres pour le groupe des futurs sexués.

Le nombre total de séquences brutes est indiqué sur le Tableau 2.2.

2.1.3 Méthodes bio-informatiques

Cette sous-partie présente les méthodes bio-informatiques utilisées pour annoter et analyser les jeux d'ARNm et de microARN.

Annotation par Blast2GO des ARNm

Blast2GO [97, 98] permet d'annoter fonctionnellement des protéines par des termes de la Gene Ontology (GO) [99]. Pour cela, Blast2GO utilise l'annotation de séquences homologues détectées à l'aide d'un BLAST [100] comparant les séquences cibles (celles que l'on souhaite annoter) avec celles d'une base de données d'intérêt. Pour une séquence cible, Blast2GO procède en trois étapes :

- Identification des séquences similaires à l'aide d'un BLASTP, paramétrable par l'utilisateur ;
- Obtention des annotations GO des séquences homologues de la base de données ;
- Association d'une annotation issue des séquences homologues à la séquence cible.

Pour la troisième étape, un score est associé à chacune des annotations qui prend en compte différents facteurs :

- La similarité entre la séquence cible et la séquence homologue ;
- L'origine de l'annotation de la séquence homologue, annotation manuelle décrivant un résultat d'expérimentation ou annotation automatique par homologie.

petits ARN		longs ARN	
banques	nombre de séquences	banques	nombre de séquences
T0	28.910.099	T0	55.889.663
	24.716.437		
	32.847.575		
T1A	29.721.245	T1A	14.949.330
	41.236.324		14.717.785
	30.541.334		15.849.753
T2A	32.101.097	T2A	25.974.062
	33.799.457		18.338.496
	32.396.234		18.354.099
T3A	37.058.813	T3A	14.456.488
	27.147.915		20.236.274
	43.652.322		1.713.8613
T1K	30.250.548	T1K	18.742.035
	29.687.320		14.711.502
	33.172.782		17.286.654
T2K	35.729.248	T2K	16.737.944
	42.780.453		16.543.737
	30.333.106		13.758.528
T3K	29.117.343	T3K	16.830.865
	25.935.181		15.506.557
	29.595.856		15.986.351

Tableau 2.2 – Nombre total de séquences par banque obtenues par extraction d'ARN sur des embryons aux stades 17, 18, 19 ou 20 (T0, T1, T2, T3) futurs sexués (A) ou asexués (K) pour les trois réplicats biologiques. Pour les petits ARN, le nombre total de séquences est la somme du nombre de séquences obtenues sur les 3 pistes de séquençage. Pour le T0 des longs ARN, le nombre total de séquences est la somme des trois réplicats de séquençage.

La ou les annotations possédant le meilleur score sont ensuite associées à la séquence cible.

Blast2GO peut aussi associer à la séquence cible une annotation commune aux séquences homologues : si les séquences homologues de la séquence cible, ou un sous ensemble de ces séquences, possèdent des termes GO avec un plus petit ancêtre commun dans la hiérarchie GO qui ne soit pas trop général (trop haut dans la hiérarchie), il pourra aussi être associé à la séquence cible.

En plus de l'annotation issue des séquences homologues obtenues par la recherche BLAST, Blast2GO annote également les séquences cibles en utilisant InterProScan [101]. InterProScan permet de détecter des domaines protéiques connus dans une séquence cible. Ces domaines connus peuvent avoir des annotations GO associées qui seront alors reportées sur la séquence cible par Blast2GO, ce qui complète ainsi l'annotation initiale.

Pour l'annotation obtenue dans la partie 2.1.1, le BLAST utilisé est un BLASTP et la base de données utilisée est la base nr[2] (version 29/07/2013) du National Center for Biotechnology Information (NCBI) avec le logiciel blast+ (version 2.2.28) [100]. Pour la recherche des domaines protéiques par InterProScan, le logiciel iprscan (version 4.8) a été utilisé contre la base de données InterPro (version 05/06/2013) [102]. La base de données de Blast2GO utilisée était la version d'août 2012. Cette base de données associe aux protéines présentes dans nr une ontologie GO si cette ontologie existe.

Identification des microARN : miRDeep2

MiRDeep2 [103] permet d'identifier des microARN matures exprimés dans un jeu de données de séquençage et leur gènes et précurseurs à l'aide d'un génome de référence. La Figure 2.2 rappelle les différentes structures associées aux microARN pour plus de clarté.

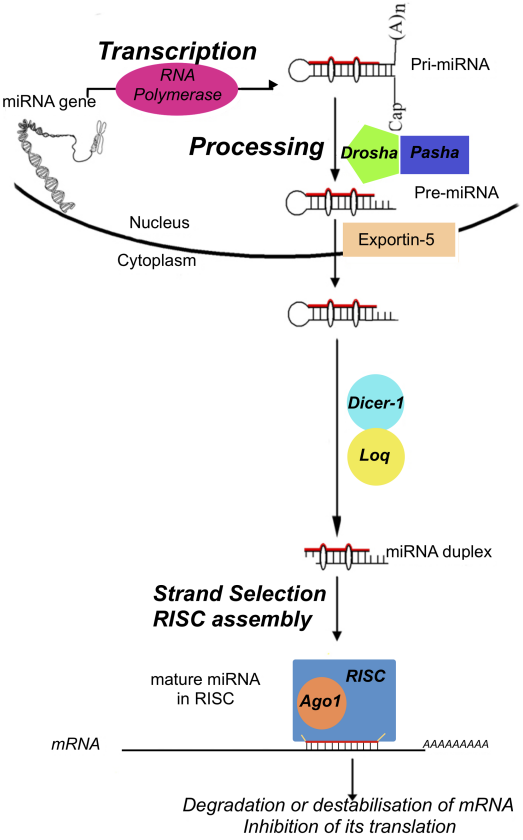


FIG. 2.2 – Rappelle des différentes structures transitoires des microARN : gène de microARN, pré-microARN, pri-microARN et microARN mature. Figure adaptée de IAGC [15].

Après nettoyage des lectures issues du séquençage, les lectures de petits ARN sont alignées sur le génome de référence et seules les séquences d'une taille d'au minimum

[2] `http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=ProgSelectionGuide`

18 nucléotides s'alignant parfaitement sur le génome sont conservées. Les séquences s'alignant parfaitement à plus de cinq positions génomiques différentes sont supprimées. Ensuite une recherche des précurseurs potentiels s'effectue pour chacun des brins du 5' vers le 3' de la façon suivante :

1. Recherche de blocs de lecture alignées avec un nombre de lecture au moins égale à 1 (hauteur). Si un autre bloc est trouvé avec une hauteur supérieure dans les 70 nucléotides suivants, alors il est choisi à la place. Cette recherche de bloc de taille maximale est faite tant qu'aucun nouveau bloc de taille supérieure n'est trouvé ;
2. Obtention de deux précurseurs potentiels issus du bloc identifié. Les précurseurs sont obtenus en extrayant le bloc ainsi que les 70 nucléotides en amont et les 20 nucléotides en aval du bloc pour le premier et les 20 nucléotides en amont et les 70 nucléotides en aval du bloc pour le second précurseur. Le bloc identifié sera soit le microARN mature 3' pour le premier précurseur, soit le microARN mature 5' pour le second (voir Figure 2.3 pour une intuition de ces extractions).

Cette extraction est répétée pour l'ensemble du génome donné en référence. Si le nombre de précurseurs potentiels identifiés est inférieur à 50.000, alors l'identification des microARN continue, sinon le seuil de taille pour détecter un bloc est augmenté de 1.

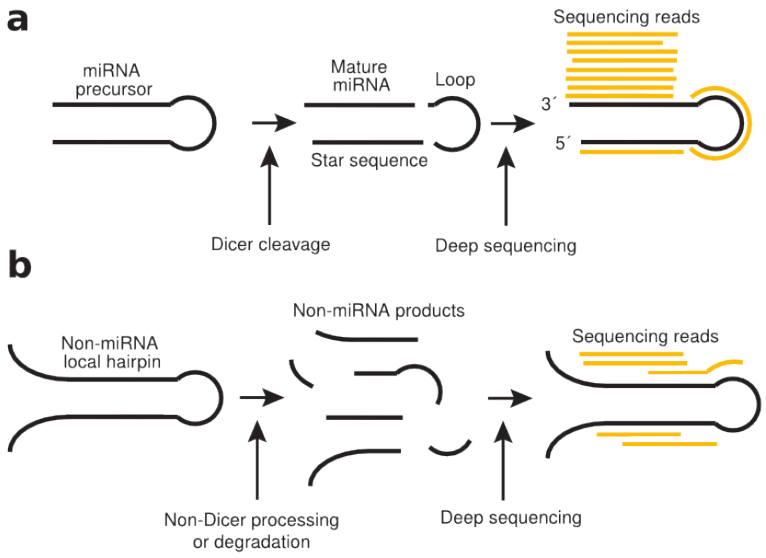


FIG. 2.3 – Les différences de séquençage et d'alignement de séquence attendues entre des tige-boucles qui sont des pré-microARN (A) et des tige-boucles qui ne sont pas des pré-microARN (B). Figure tirée de Friedlander *et al.* [104].

Une fois les précurseurs obtenus, l'ensemble des lectures est réaligné sur les précurseurs potentiels, cette fois-ci en autorisant au maximum un mésappariement par alignement. Si une lecture s'aligne parfaitement sur un précurseur potentiel, alors seul l'alignement parfait est gardé. Cet alignement est aussi mené sur les microARN matures déjà identifiés pour cette espèce s'ils sont fournis mais sans mésappariement autorisé. À noter que seuls les alignements sur le même brin que le précurseur sont gardés.

L'étape suivante est la prédiction des structures secondaires des précurseurs potentiels par RNAfold [105]. Une fois la structure secondaire la plus probable obtenue pour chacun des précurseurs potentiels, les séquences des microARN matures et de la boucle sont définies de la façon suivante : un des microARN mature est défini en fonction du bloc qui a servi à identifier ce précurseur, le second microARN mature est la séquence complémentaire du premier en fonction de la structure secondaire obtenue et la boucle comme la séquence entre ces deux séquences. De plus, les lectures qui ne s'alignent pas de façon cohérente avec le clivage du précurseur par Dicer sont supprimées, c'est-à-dire que les lectures doivent s'aligner uniquement sur l'un des deux brins ou sur la boucle (voir Figure 2.3) avec une tolérance de deux nucléotides. Des filtres supplémentaires sont ensuite mis en place :

- La tige-boucle formée ne doit pas avoir de bifurcation ;
- 60 % ou plus des nucléotides du microARN mature défini par le bloc doivent être appariés ;
- Moins de 10 % des lectures doivent ne pas être alignées de façon cohérente avec le clivage par Dicer.

La probabilité pour une séquence d'ARN donnée que l'énergie libre minimum de sa structure secondaire soit différente de celle d'une séquence aléatoire est estimée à l'aide du logiciel randfold [106]. Puis, pour chacun des précurseurs, un score est calculé qui représente la probabilité que le précurseur potentiel soit un vrai pré-microARN. Pour plus de détail sur le calcul de ce score, voir la publication de miRDeep par Friedlander *et al.* [104].

Afin de prédire les microARN chez le puceron du pois, les séquences de petits ARN obtenues sur les embryons asexués et sexués ont été utilisées. MiRDeep2 a été utilisé sur ces lectures avec le génome de référence d'*A. pisum* et les microARN matures précédemment identifiés par Legeai *et al.* [81].

Regroupement des gènes de microARN : l'outil cluster de la suite logicielle bedtools

BedTools est une suite logicielle développée par Quinlan et Hall [107] permettant d'effectuer un grand nombre d'analyses génomiques. Notamment, nous avons utilisé l'outil *cluster* qui permet de regrouper des éléments génomiques en fonction de leurs positions sur le génome. Cet outil regroupe les gènes au sein du même cluster si, de proche en proche, ils sont à une distance qui est inférieure à un seuil fixé par l'utilisateur. Mathématiquement, c'est une approche qui permet d'obtenir les composantes connexes du graphe des éléments à une distance inférieur au seuil. Un autre paramètre permet de choisir si le regroupement est effectué pour des gènes de microARN qui sont sur le même brin ou non.

Ce logiciel a été utilisé pour définir les clusters des gènes de microARN sur un même brin et avec une distance seuil de 2 kb à la fois pour *Acyrthosiphon pisum* et pour *Drosophila melanogaster*.

Prédiction des sites de fixation des microARN matures : TargetScan version 5

La méthode TargetScan [51] permet de prédire des sites de fixation de microARN matures sur des séquences, classiquement les 3'UTR d'ARNm, puis de leur associer un score qui représente l'efficacité de la répression du microARN mature sur l'ARNm. Pour ce faire, la méthode procède d'abord à la détection des sites de fixation potentiels par complémentarité de séquence entre une sous séquence du microARN mature et la séquence du 3'UTR de l'ARNm.

Trois types de complémentarité sont considérés comme des sites de fixation potentiels décris ci-dessous et présentés Figure 2.4 :

- Une complémentarité du 2ème au 7ème nucléotide du microARN mature sur le 3'UTR de l'ARNm et une adénine directement en aval du site de fixation sur le 3'UTR appelé 7mer-A1 ;
- Une complémentarité du 2ème au 8ème nucléotide du microARN mature sur le 3'UTR de l'ARNm appelé 7mer-m8 ;
- Une complémentarité du 2ème au 8ème nucléotide du microARN mature sur le 3'UTR de l'ARNm et une adénine directement en aval du site de fixation sur le 3'UTR appelé 8mer.

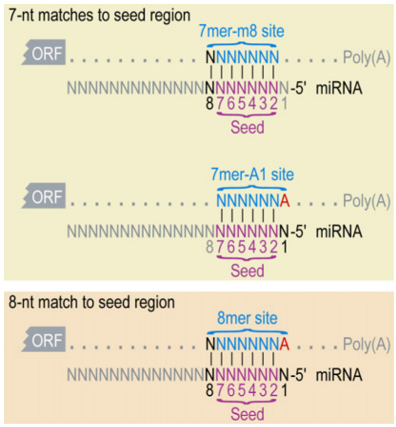


FIG. 2.4 – Les différents sites de fixation utilisés par TargetScan. Image adaptée de Grimson *et al.* [51].

Les auteurs de la méthode ont défini un modèle afin de calculer, pour chacun des sites de fixation détectés, un score global à partir de 4 critères :

- Le type de site de fixation : l'efficacité de la répression est la plus élevée pour les sites de type 8mer, puis 7mer-m8 et 7mer-A1 a la répression la plus faible ;
- La position du site de fixation sur le 3'UTR : la répression est plus élevée s'il se situe à l'une des extrémités ;
- Le pourcentage en A/U autour du site de fixation : plus ce pourcentage est élevé plus la répression est forte ;
- La complémentarité du 3' du microARN mature avec le 3'UTR : la répression est d'autant plus élevée que la complémentarité est forte.

Le score global obtenu reflète l'efficacité de la fixation du microARN mature sur l'ARNm à l'aide de ce site de fixation. À noter que plus le score global est bas, plus l'efficacité

prédite est élevée.

De plus, si un site de fixation est à 10 nucléotides ou moins du codon stop (le début du 3'UTR) alors TargetScan ne calcule pas de score global. Il indique que le site de fixation est trop proche par rapport à l'encombrement stérique du complexe protéique RISC (microARN mature et protéines argonautes) et de la machinerie de la traduction.

Afin de prédire les sites de fixation chez *Acyrthosiphon pisum*, l'ensemble des 3'UTR disponibles (35.462) et des microARN matures (802) ont été utilisés avec la version 5.0 de TargetScan. TargetScan prédit un ensemble de 4.230.168 sites de fixation dont 90.410 sont notés comme trop proches (voir ci-dessus). Les 4.139.758 sites de fixation restants sont utilisés pour la suite des analyses.

Il est à noter que TargetScan permet de ne garder que les complémentarités entre microARN matures et 3'UTR qui sont conservées chez plusieurs espèces. Pour ce faire, TargetScan utilise les alignements multiples des séquences des microARN matures et des 3'UTR pour définir les éléments conservés. Nous n'avons pas pu utiliser cette possibilité car nous n'avions pas de prédictions de microARN matures et de 3'UTR pour des espèces de puceron autres qu'*Acyrthosiphon pisum*.

2.1.4 Bases de données utilisées

MiRBase

MiRBase [19, 20, 21, 22, 23] est une base de données répertoriant des séquences de microARN publiées, annotées et mises à jour par la communauté scientifique. Ces annotations comportent à la fois les microARN matures, les précurseurs de microARN avec leur structure secondaire, les positions correspondant aux précurseurs de microARN et si disponible, le nombre de banques de séquençage, le nombre de lectures correspondant aux différents microARN matures et les cibles des microARN matures prédites par différentes méthodes. En plus de ces annotations, miRBase fournit des règles pour nommer les nouveaux microARN identifiés. La version de la base de données que nous avons utilisée était la version 20 datant de Juin 2013. Une version plus récente existe (Juin 2014) mais les résultats ont été obtenus avant cette mise à jour.

RefSeq

RefSeq [108] est une base de données du NCBI de séquences non-redondantes et annotées par les membres du NCBI. Suivant les séquences, ces annotations sont à la fois obtenues par un processus d'annotation automatique et par une annotation experte.

AphidBase

AphidBase [78], développée en partie dans l'équipe Écologie et Génétique des Insectes, regroupe un ensemble de ressources disponibles pour *Acyrthosiphon pisum* telles que l'annotation des gènes, l'annotation des protéines ou encore la visualisation du génome d'*Acyrthosiphon pisum* et des gènes.

PhylomeDB

La base de données PhylomeDB version 4 [109] regroupe les arbres phylogénétiques de gènes d'ARNm permettant de définir des relations d'orthologie et de paralogie. Ainsi, il est possible de savoir si deux gènes sont issus de la duplication d'un même gène ancestral pour différentes espèces.

FlyBase

FlyBase [110] est une base de données qui regroupe les génomes, les gènes et leurs annotations pour différentes espèces de *Drosophila*. Elle permet notamment de télécharger les séquences des gènes, des transcrits ou encore des 3'UTR des gènes.

2.2 Résultats : catalogue et caractérisation des ARNm, des microARN et de leurs interactions

Cette partie présente l'extraction des 3'UTR des ARNm, l'analyse du nouveau catalogue des gènes codant pour des microARN et les interactions entre ces ARNm et ces microARN pour le puceron du pois *Acyrthosiphon pisum*. Ces listes ont été établies en se basant sur la version 2.1 du génome d'*A. pisum* LSR1 ainsi que sur des données de séquençage décrites dans la partie 2.1.2.

2.2.1 Les ARNm

Prédiction des 3'UTR

Les microARN reconnaissant de façon préférentielle les régions 3'UTR des ARNm, Fabrice Legeai a réalisé l'inventaire de ces régions. Deux cas de prédiction de 3'UTR se présentent pour un ARNm : soit le 3'UTR peut être prédit en intégrant des lectures issues de RNA-Seq s'alignant en aval du codon stop prédit, soit les 1.000 bases en aval du codon stop prédit ont été définies comme le 3'UTR prédit. Sur l'ensemble des 36.990 ARNm, 1.528 n'ont pas de 3'UTR car ils sont à la limite d'un scaffold.

2.2.2 Les microARN

Définitions

Nous distinguerons un gène de microARN, un précurseur de microARN et un microARN mature comme suit :

- *Gène de microARN* : c'est la position de début et de fin sur le génome de la séquence codant pour un précurseur de microARN ;
- *Précurseur de microARN* : c'est la séquence d'ARN qui est transcrite à partir du gène du microARN et qui se replie en une structure en tige-boucle ;
- *MicroARN matures 5p et 3p* : ce sont les microARN qui vont se fixer sur le 3'UTR de l'ARNm et potentiellement réprimer sa traduction ou induire son clivage. Ils sont tous deux issus de la tige du précurseur du microARN. Les microARN matures 5p et 3p sont respectivement situés en amont et en aval de la boucle du précurseur.

Deux gènes différents (à des positions différentes sur le génome) peuvent présenter la même séquence et donc produire le même précurseur.

Les noms des microARN sont constitués de plusieurs parties, toutes séparées par un « - », comme par exemple *api-mir-1-5p* :

- La première partie représente l'espèce, ici « *api* » pour *Acyrthosiphon pisum* ;
- La deuxième partie est soit « *mir* » pour un précurseur soit « *miR* » pour un microARN mature ;
- La troisième partie représente le numéro du précurseur de microARN associé. Si la séquence du précurseur a déjà été identifiée dans une autre espèce (donc est présente dans la base de données miRBase), alors le numéro sera identique à celui de l'autre espèce. Sinon ce numéro est incrémental à partir d'un numéro qui n'est pas présent dans la base miRBase ;

- La quatrième partie est spécifique aux microARN matures. Les noms seront suivis par « *-5p* » et « *-3p* » respectivement pour les microARN matures 5p et 3p.

À noter que ici, les gènes de microARN qui sont nouvellement identifiés par les nouveaux jeux de séquençage chez *A. pisum* porteront comme identifiant « *novelX* » avec X allant de 1 au nombre maximum de nouveaux gènes identifiés. Par exemple, api-mir-novel2 est le deuxième nouveau gène de microARN identifié chez *A. pisum* avec les nouveaux jeux de données. Il a été décidé de prendre cette notation car cela permet de ne décider de la numérotation finale uniquement au moment de son intégration effective de miRBase. À noter que les gènes précédemment identifiés et qui sont spécifiques à *A. pisum* portent un identifiant supérieur à 3000 donné par miRBase.

Il existe des cas particuliers où des précurseurs de microARN différents produisent des microARN matures avec des séquences identiques ou proches. MiRBase [23] fournit un ensemble de règles pour discriminer ces différents cas :

- Si des précurseurs sont identiques et que les gènes associés sont à des positions différentes ou que les précurseurs sont différents mais que les microARN matures ont des séquences strictement identiques, alors les précurseurs auront le même numéro mais seront différenciés par l'ajout en fin de nom d'un « - » suivi par un numéro ;
- Si des précurseurs différents ont des microARN matures avec des séquences proches, alors les précurseurs porteront le même numéro mais seront différenciés par une lettre en fin de nom.

Par exemple, les précurseurs api-mir-3051-1 et api-mir-3051-2 ont des séquences différentes mais les microARN matures api-mir-3051-1-5p et api-mir-3051-2-5p ont des séquences identiques, de même pour api-mir-3051-1-3p et api-mir-3051-2-3p. Les précurseurs api-mir-263a et api-mir-263b ont des séquences différentes et les séquences de api-mir-263a-5p et api-mir-263b-5p sont proches, de même pour api-mir-263a-3p et api-mir-263b-3p. Les précurseurs de microARN avec le même numéro sont définis comme étant de la même famille. Ici api-mir-3051-1 et api-mir-3051-2 font partie de la même famille.

Pour définir si deux séquences de microARN matures sont proches, aucune règle stricte n'étant fournie par miRBase, nous avons défini les règles suivantes. Deux microARN matures différents ont des séquences dites proches si :

- Leurs nucléotides aux positions allant de 2 à 7 sont identiques, positions qui correspond classiquement à la définition de la graine du microARN ;
- Sur le reste des microARN matures, seulement deux délétions, insertions et/ou mésappariements existent au maximum après alignement des microARN matures.

Ces règles ont été définies pour avoir une notion « fonctionnelle » dans les familles, c'est à dire que deux microARN matures issus d'une même famille cibleront potentiellement les mêmes ARNm. À noter que, contrairement à la convention de miRBase, il a été décidé de se référer aux microARN matures par la notation « mir » et non pas « miR ». La notation avec et sans majuscule est utilisée pour ne pas confondre le pré-microARN et le microARN mature, mais ici nous nous référerons au microARN mature en ajoutant toujours « -5p » ou « -3p » ce qui empêche la confusion entre précurseur et mature.

Identification des gènes, précurseurs et molécules matures de microARN et de leurs familles

Pour identifier les gènes de microARN, le logiciel miRDeep2 [103] a été utilisé par Fabrice Legeai. Pour ce travail, l'annotation des gènes de microARN d'*Acyrthosiphon pisum* s'est basée sur les séquences des petits ARN des 63 banques obtenues au laboratoire (voir 2.1.2).

À partir de ce jeu de données, miRDeep2 prédit un ensemble de 445 gènes codant pour des microARN. Seulement 401 des précurseurs associés à ces gènes possèdent une structure secondaire conforme avec la structure secondaire en tige-boucle des pré-microARN, c'est-à-dire avec une probabilité obtenue par randfold inférieure à 0,05. Il a donc été décidé de ne garder que ces 401 gènes de microARN pour la suite de l'analyse. Ces 401 gènes correspondent à 329 séquences de précurseurs et produisent 288 séquences uniques de microARN matures 5p et 285 séquences uniques de microARN matures 3p (573 séquences de microARN matures au total). Sur l'ensemble de ces 401 gènes, 39 ($\sim$10 %) sont des gènes déjà connus pour d'autres espèces et 362 sont des gènes de microARN identifiés spécifiquement chez *A. pisum*. Sur ces 362 gènes, 40 ($\sim$11 %) étaient déjà identifiés dans Legeai *et al.* [81] et 322 ont été identifiés à l'aide de ces nouveaux jeux de séquençage. Le pourcentage peu élevé de gènes de microARN identifiés chez d'autres espèces peut s'expliquer par les critères plus stricts utilisés ici pour définir la similarité entre deux gènes/précurseurs de microARN que ceux potentiellement utilisés par miRBase. À noter que les microARN api-mir-let-7 et api-mir-bantam sont retrouvés, même avec nos critères plus stricts.

Parmi les espèces du phylum Hexapoda présentes dans la base de données miRBase (ver. 20) [23] et en se basant sur le dernier catalogue, *A. pisum* se place parmi les espèces possédant un nombre important de microARN matures uniques. Le Tableau 2.3 montre le nombre de séquences uniques de précurseurs de microARN et de microARN matures 5p et 3p pour *Acyrthosiphon pisum* et certaines espèces du phylum du puceron du pois (Hexapoda) issue de miRBase.

espèces	nombre de précurseur	nombre de microARN matures
Acyrthosiphon pisum	401	573
Bombyx mori	489	567
Tribolium castaneum	220	430
Drosophila melanogaster	238	426
Drosophila pseudoobscura	210	273
Aedes aegypti	101	124
Anopheles gambiae	67	65

Tableau 2.3 – Nombre de séquences uniques de précurseurs et de microARN matures 5p et 3p pour plusieurs espèces du phylum Hexapoda. Données issues de miRBase [23].

Au sein du phylum Hexapoda, les nombres de séquences uniques pour les précurseurs de microARN et pour les microARN matures sont variables, ce qui peut refléter soit des niveaux d'annotation des microARN de qualités différentes dans les espèces citées, soit des différences biologiques encore non expliquées. Cependant, on peut faire

l'hypothèse que le génome de *D. melanogaster* est bien annoté alors qu'il présente un nombre inférieur de microARN comparé au puceron du pois. Cette différence pourrait s'expliquer par un niveau de duplication des gènes chez *A. pisum* plus élevé (voir plus loin partie 2.2.2).

Familles des gènes de microARN Sur les gènes de microARN d'*A. pisum*, 200 gènes sur les 401 (50 %) se répartissent sur 66 familles avec au moins deux gènes. La Figure 2.4 présente la répartition des 200 gènes dans les 66 familles. Sur ces 66 familles, 63 (95 %) sont constituées de microARN identifiés uniquement chez *A. pisum*. Les trois familles avec des gènes connus sont les familles api-mir-2, api-mir-92 et api-mir-263 toutes trois constituées de deux gènes de microARN. Ces trois familles existent aussi chez d'autres espèces, comme *D. melanogaster*. Le Tableau présente aussi la répartition en familles pour les gènes de microARN chez *D. melanogaster*, basée sur les microARN annotés dans miRBase. On peut voir que la proportion de gènes au sein d'une famille chez *D. melanogaster* est largement inférieure à celle chez *A. pisum* (30 gènes sur 238 (∼13 %) contre 50 %), et que les nombres de familles diffèrent (12 contre 66). De plus la diversité dans les tailles des familles chez *A. pisum* est supérieure à celle de *D. melanogaster*. Les différences en nombre de familles, en nombre de gènes impliqués dans des familles et sur la taille des différentes familles pourrait être expliquées par un duplication importante des gènes de microARN après la spéciation entre les deux branches évolutives correspondant à *A. pisum* et *D. melanogaster*. Deux faits supportent cette hypothèse. D'une part 95 % des familles sont constituées de gènes identifiés pour l'instant uniquement chez le puceron du pois, et d'autre part il a été montré qu'un grand nombre de gènes codant pour des protéines ont été dupliqués chez le puceron [15]. On peut donc faire l'hypothèse que cette duplication a aussi eu lieu pour les gènes de microARN.

espèce	*Acyrthosiphon pisum*		*Drosophila melanogaster*	
taille	nombre de familles	nombre de gènes	nombre de familles	nombre de gènes
2	30	60	8	16
3	20	60	3	9
4	7	28	0	0
5	4	20	1	5
6	4	24	0	0
8	1	8	0	0
total	66	200	12	30

Tableau 2.4 – Répartition des gènes de microARN appartenant à des familles d'au moins deux gènes chez *Acyrthosiphon pisum* et *D. melanogaster*.

Localisation génomique des gènes de microARN

Classification des positions de gènes de microARN On rappelle ici que les positions génomiques des gènes de microARN peuvent être séparées en deux classes principales :

- *Intergénique* : le gène du microARN est situé entre deux gènes d'ARNm ;
- *Intragénique* : le gène du microARN est situé au sein d'un gène d'ARNm. Par la suite on appellera « *ARNm hôte* » ou « *gène hôte* » des ARNm ou gènes contenant un ou plusieurs gènes de microARN.

Un gène de microARN intragénique sera sous le même contrôle transcriptionnel que son gène hôte. Un gène de microARN intergénique sera sous le contrôle transcriptionnel de sa propre séquence promotrice. Pour les gènes de microARN intragéniques, on peut s'attendre à ce que ces gènes de microARN et leur gène d'ARNm hôte soient co-exprimés car ils seront transcrits ensemble.

On distingue deux sous-classes de gènes de microARN intragéniques :

- *Intronique* : le gène du microARN est situé dans un intron du gène hôte ;
- *Exonique* : le gène est situé dans un exon du gène hôte.

Le Tableau 2.5 résume la classification intergénique, intronique et exonique des 401 gènes de microARN pour *Acyrthosiphon pisum*.

position du gène du microARN	nombre de gènes de microARN
intergénique	253
intronique	102
exonique	46
total	401

Tableau 2.5 – Classification des 401 gènes de microARN intergéniques, introniques ou exoniques, chez *A. pisum*.

Sur les 401 gènes de microARN, 63 % des gènes sont intergéniques, 25 % des gènes sont introniques et 12 % des gènes sont exoniques. Comme attendu, la majorité des gènes est localisée au niveau intergénique ou intronique (88 %). La répartition des localisations observée est en accord avec la répartition de l'ancienne annotation des microARN obtenue sur le puceron du pois avant le début de ce travail [81]. Néanmoins, la proportion de gènes de microARN exoniques est supérieure à celle observée chez d'autres espèces [36, 111, 112, 113].

Clusters de microARN Les gènes de microARN sont parfois regroupés en clusters dans un même environnement génomique. On appelle cluster génomique de microARN un ensemble de gènes de microARN qui se suivent sur le génome à des positions génomiques proches. Les gènes de microARN appartenant au même cluster sont des gènes qui sont potentiellement sous le contrôle du même promoteur et qui peuvent être transcrits au sein du même pri-microARN polycistronique [114, 115]. Les microARN matures provenant de ces différents gènes clusterisés sont donc potentiellement co-exprimés.

Pour annoter ces clusters de gènes de microARN du puceron du pois, l'outil *cluster* de la suite d'outils *bedtools* [107] a été utilisé (partie 2.1.3 pour plus de détails). Une distance seuil maximale entre deux gènes consécutifs de microARN de 2kb a été utilisée et les gènes de microARN ont de plus été regroupés uniquement s'ils étaient sur le même brin génomique. Le Tableau 2.6 présente le nombre de clusters de microARN en fonction du nombre de gènes de microARN présents dans les clusters obtenus avec l'outil cluster.

Sur l'ensemble des 401 gènes, 207 gènes de microARN (51,6 %) sont répartis dans 52 clusters. Ces clusters ont une taille génomique moyenne de 2,2 kb et sont constitués

taille des clusters en nombre de gènes	nombre de clusters
2 gènes	21
3 gènes	10
4 gènes	5
5 gènes	6
6 gènes	6
7 gènes	1
12 gènes	1
14 gènes	1
16 gènes	1
total	52

Tableau 2.6 – Nombre de clusters de microARN en fonction du nombre de gènes de microARN présents dans les clusters pour *Acyrthosiphon pisum*. Il faut noter que certains clusters peuvent ne pas avoir été identifiés ou être incomplets si un gène ou un cluster est à l'une des extrémités d'un scaffold.

en moyenne de 4 gènes. La majorité des clusters regroupent entre 2 et 6 gènes avec plus de la moitié des clusters qui incluent 2 ou 3 gènes. Le nombre de gènes de microARN en cluster est supérieur à celui trouvé chez d'autres espèces animales où le pourcentage de gènes de microARN en cluster est au maximum de 40 % [116].

Afin de comparer les clusters obtenus sur *Acyrthosiphon pisum* avec ceux d'une autre espèce, le même protocole a été appliqué à *Drosophila melanogaster*. La Figure 2.5 présente l'histogramme comparatif entre le nombre de clusters en fonction du nombre de gènes de microARN au sein du cluster pour *A. pisum* et *Drosophila melanogaster* en appliquant le même protocole (avec les données de miRBase [23]).

Sur l'ensemble des 238 gènes de microARN de *Drosophila melanogaster*, la répartion des gènes en clusters est différente avec 70,6 % qui sont en singleton (pas dans un cluster) et 29,4 % repartis dans un total de 22 clusters. De la même façon que pour le puceron du pois, plus de la moitié des clusters regroupent 2 ou 3 gènes de microARN. Néanmoins, les clusters de taille plus importante sont en plus petit nombre avec seulement 1 cluster pour chacune des tailles de 4 gènes, 5 gènes et 6 gènes pour *Drosophila melanogaster* contre 6 clusters en moyenne pour les mêmes tailles de gènes chez *Acyrthosiphon pisum*. De plus, il n'y a pas de cluster de « grande taille » (clusters de taille 12, 14, et 16 Tableau 2.6 et Figure 2.5). Afin d'observer les spécificités de ces 3 clusters de « grande taille » du puceron du pois, les alignements multiples des séquences des gènes de microARN contenus dans ces clusters sont présentés Figure 2.6, 2.7 et 2.8 respectivement pour les clusters de taille 12, 14 et 16. Les alignements ont été obtenus en utilisant ClustalW [117, 118].

Pour les gènes de microARN du cluster de taille 12, on peut voir la présence de deux familles complètes, api-mir-novel14 (quatre gènes) et api-mir-novel10 (deux gènes) et que de plus les séquences entre ces familles sont très proches. api-mir-novel188 et api-mir-novel117 ont eux des séquences très proches de la famille api-mir-novel14 et api-mir-novel202 a lui une séquence qui s'approche de celle de la famille api-mir-novel10. Les trois autres gènes n'ont pas de grande similarité avec les autres séquences. À noter que le gène api-mir-3032b fait partie d'une famille où il y a un autre gène, api-mir-3032a

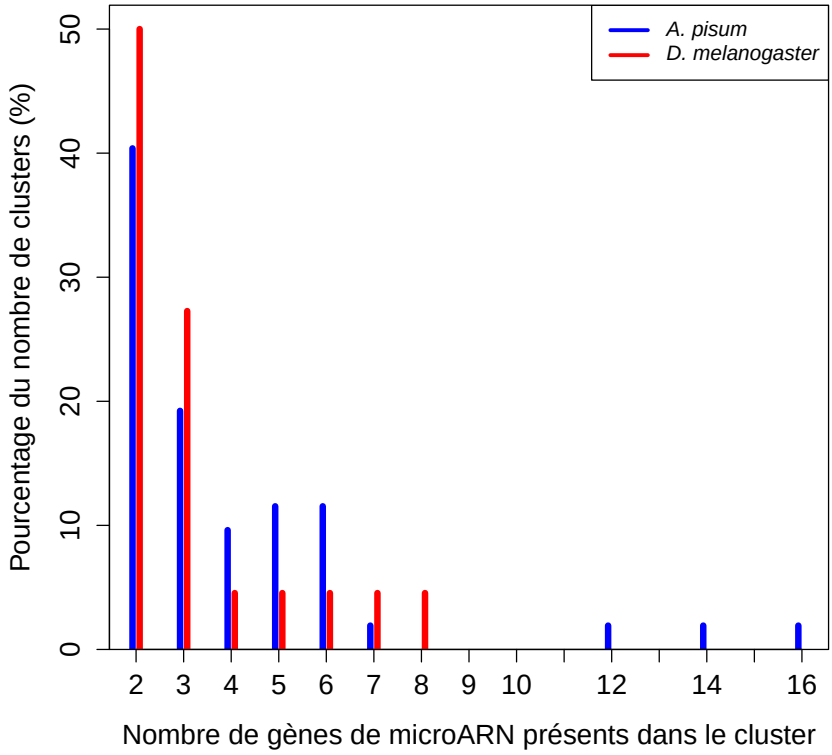


FIG. 2.5 – Histogramme du pourcentage du nombre de clusters en fonction du nombre de gènes de microARN au sein des clusters. En bleu le nombre de clusters pour *A. pisum* et en rouge le nombre de clusters pour *D. melanogaster*.

mais que ce gène est sur un autre scaffold.

Sur le cluster de taille 14, trois familles sont complètes : api-mir-novel38 (trois gènes), api-mir-novel16 (trois gènes) et api-mir-novel33 (quatre gènes) avec une très grande similarité de séquence entre ces trois familles. Le précurseur api-mir-novel58 n'a que deux substitutions, présentes dans le mature 3p, avec ceux de la famille api-mir-novel33. Le précurseur api-mir-novel29a-2, lui aussi avec une séquence proche des précurseurs précédents, est le seul représentant de sa famille, qui est constituée de quatre gènes. Le seul autre gène de cette famille aussi présent sur ce cluster est api-mir-novel29b-2, mais ce gène se situe à une distance de 3.199 nucléotides de la fin du cluster de taille 14, distance supérieure au seuil utilisé pour obtenir ces clusters (pour rappel 2.000 nucléotides). Les deux derniers microARN, api-mir-novel21a-1 et api-mir-novel21a-4 font partie de la même famille qui est constituée de six gènes. Les quatre autres gènes de la famille sont présents sur le même scaffold mais à une distance de 6.784 nucléotides du cluster.

Le dernier cluster de grande taille, celui de 16 gènes, n'est constitué que de familles complètes : api-mir-3055 (cinq gènes), api-mir-novel41 (quatre gènes), api-mir-novel43 (six gènes). Il n'y a que api-mir-novel195, qui n'appartient à aucune famille. Les sé-

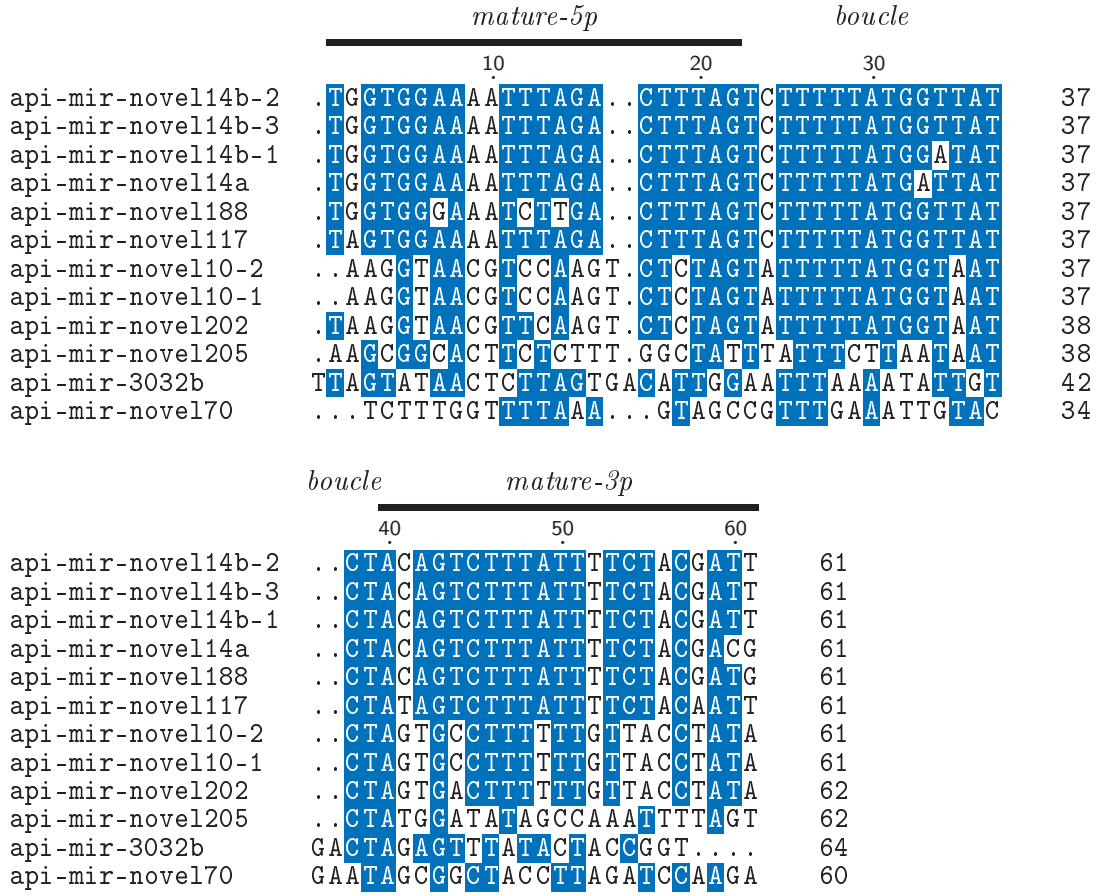


FIG. 2.6 – Alignement multiple des précurseurs du cluster de taille 12. Les nucléotides présents sur plus de 50 % des séquences sont représentés sur fond bleu, les autres sur fond blanc ; les positions des microARN matures 5p et 3p et de la boucle, définies sur api-mir-novel14b-2, sont schématisées en traits gras. Les alignements ont été obtenus avec ClustalW. Figure obtenue à l'aide du paquet LaTeX TeXshade [119].

quences sont très proches avec aucune insertion ou délétion et très peu de substitutions dans l'alignement.

Malgré la « grande taille » de ces trois clusters en comparaison des clusters obtenus chez *D. melanogaster*, la très forte similarité de l'ensemble des gènes de microARN contenus au sein de chacun de ces clusters laisse supposer que leur apparition est due à la duplication multiple d'un même gène ancestral pour chacun de ces clusters.

Classification des positions des clusters de gènes de microARN par rapport aux gènes d'ARNm Comme pour les gènes de microARN, les positions des clusters peuvent être classifiées en plusieurs catégories selon qu'ils apparaissent à l'intérieur de gènes d'ARNm, exon ou intron, ou entre deux gènes d'ARNm. Les positions des clusters dépendent des positions des microARN qui sont au sein de ces clusters. À un cluster

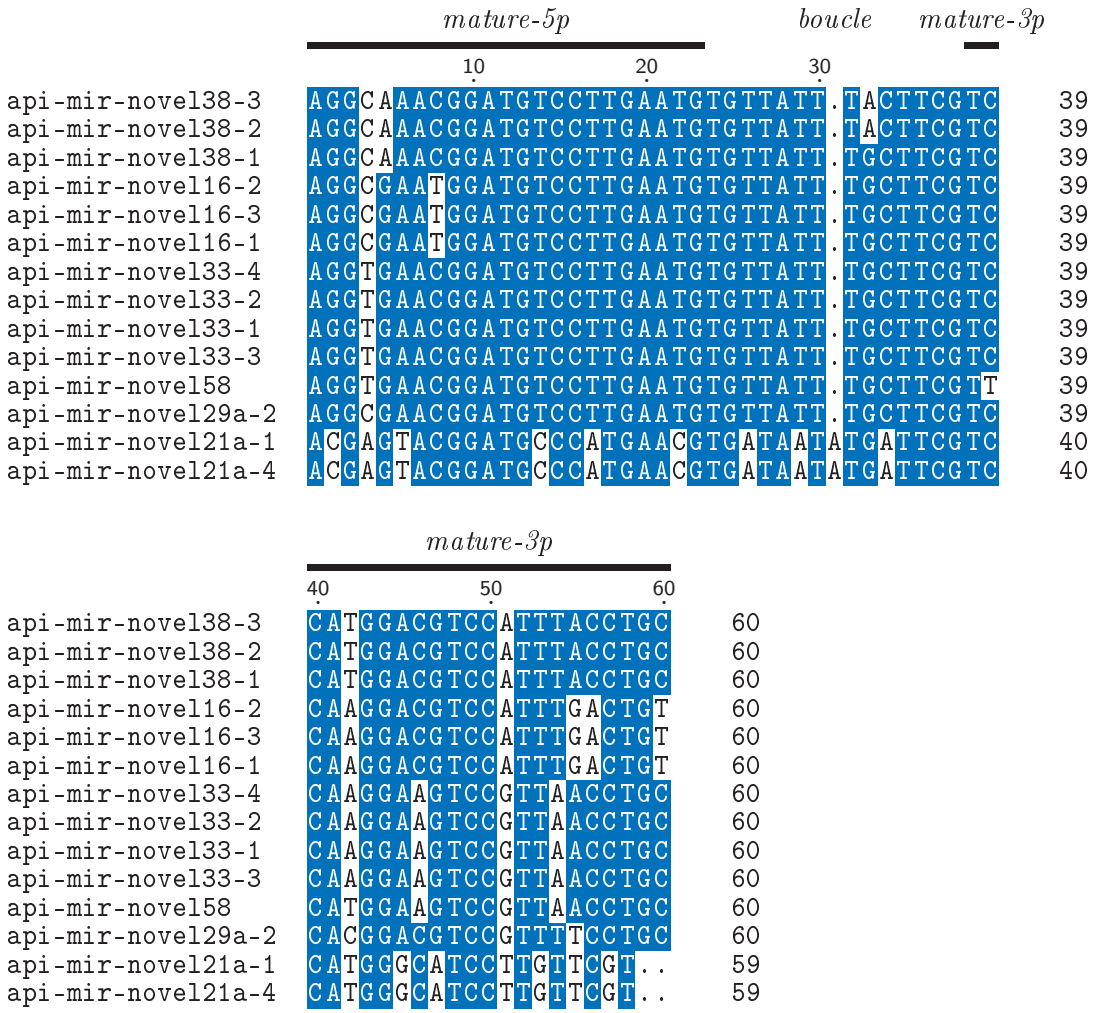


FIG. 2.7 – Alignement multiple des précurseurs du cluster de taille 14. Les nucléotides présents sur plus de 50 % des séquences sont représentés sur fond bleu, les autres sur fond blanc ; les positions des microARN matures 5p et 3p et de la boucle, définies sur api-mir-novel38-3, sont schématisées en traits gras. Les alignements ont été obtenus avec ClustalW. Figure obtenue à l'aide du paquet LaTeX TeXshade [119].

de gènes on associe l'ensemble des classes de chacun de ses éléments. Le Tableau 2.7 montre les effectifs des différents ensembles observés.

La classification des positions des clusters de gènes de microARN suit globalement la distribution de la classification des gènes avec 71 % des clusters qui sont intergéniques, 12 % sont introniques et 5 % sont exoniques. Sur l'ensemble des 52 clusters, 6 possèdent des positionnements multiples (12 %). Cette classification multiple pourrait être la conséquence d'une mauvaise annotation des gènes de microARN ou d'ARNm. La description de 3 de ces 6 clusters est donc présentée ci-dessous.

Le cluster contenant les gènes api-mir-novel18-2, api-mir-novel18-3 et api-mir-novel42-

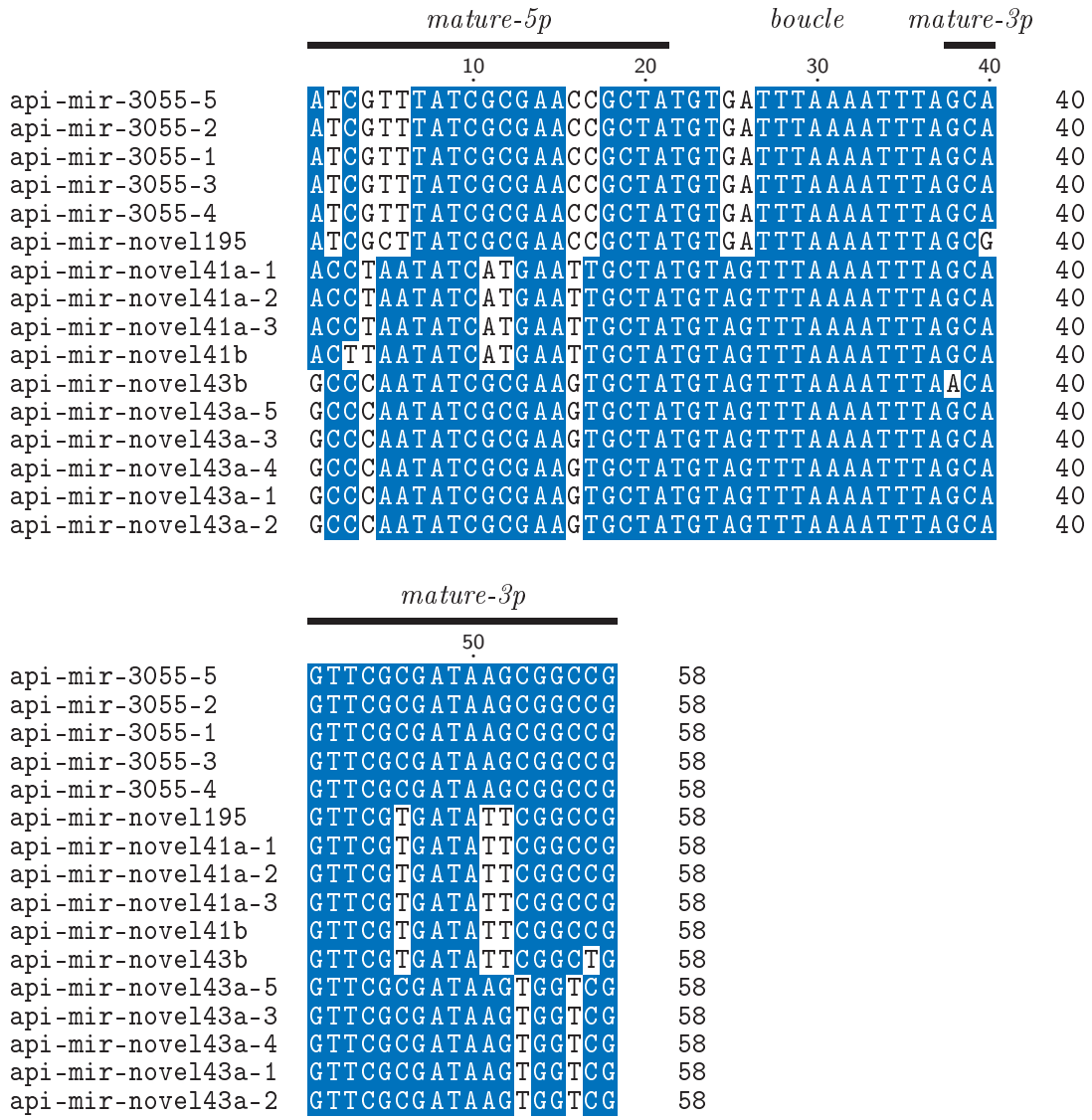


FIG. 2.8 – Alignement multiple des précurseurs du cluster de taille 16. Les nucléotides présents sur plus de 50 % des séquences sont représentés sur fond bleu, les autres sur fond blanc ; les positions des microARN matures 5p et 3p et de la boucle, définies sur api-mir-3055-5, sont schématisées en traits gras. Les alignements ont été obtenus avec ClustalW. Figure obtenue à l'aide du paquet LaTeX TeXshade [119].

2 (exonique et intergénique, Figure 2.9) possède deux gènes de microARN exoniques issus de la même famille api-mir-novel18 et un gène intergénique api-mir-novel42-2. La distance entre ces 3 gènes de microARN est de 1.240 nucléotides entre api-mir-novel18-3 et api-mir-novel18-2 et de 497 nucléotides entre api-mir-novel18-2 et api-mir-novel42-2. Les deux gènes exoniques sont au sein du même exon du 5'UTR du gène hôte ACYPI54131, dont un est situé à la quasi fin de cet exon. Le gène intergénique est lui

position du cluster	nombre de clusters
{intergénique}	37
{intronique}	6
{exonique}	3
{exonique, intergénique}	3
{intronique, intergénique}	2
{exonique, intronique}	1
total	52

Tableau 2.7 – Classification des positions génomiques des clusters de microARN et leur effectifs.

situé directement en aval du début du gène ACYPI54131. Concernant les séquences de ces trois gènes, api-mir-novel42-2 possède une séquence très proche des deux autres avec seulement trois substitutions aux nucléotides 20 (A->G), 45 (A->T) et 58 (A->G). La similarité de ces séquences laisse supposer que ces trois gènes sont issus du même gène de microARN ancestral et que l'annotation de ce cluster est correcte.

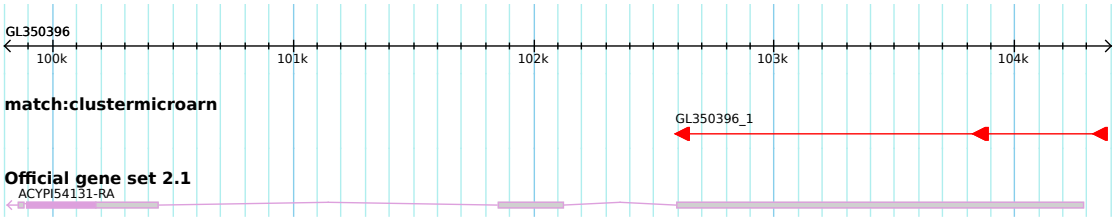


Fig. 2.9 – Région génomique contenant le cluster GL350396_1 des microARN api-mir-novel18-3, api-mir-novel18-2 et api-mir-novel42-2 et le gène hôte ACYPI54131. Le cluster de microARN est représenté en haut en rouge. Les flèches schématisent les gènes de microARN et le brin. Les gènes sont représentés en bas en violet. Les rectangles schématisent les exons (foncé pour les régions traduites, clair pour les régions non traduites), les traits les introns et la flèche le brin. La figure a été extraite d'AphidBase.

Le deuxième cluster (exonique et intronique, Figure 2.10) contient tous les gènes de la famille api-mir-3043 : api-mir-3043-1, api-mir-3043-2 et api-mir-3043-3. Deux gènes sont exoniques (api-mir-3043-1 et api-mir-3043-3) et l'autre est intronique. Ces 3 gènes sont quasiment équidistants les uns des autres avec des distances de 1.341 nucléotides et 1.336 nucléotides. Ces 3 gènes de microARN sont inclus dans 3 gènes hôtes ACYPI48456, ACYPI082069 et ACYPI080956. Les deux gènes hôtes ACYPI48456 et ACYPI082069 possèdent une annotation RefSeq [108] mais aucun ne possède une protéine homologue connue. Pour le gène hôte ACYPI080956, aucune annotation RefSeq ou protéine homologue n'est connue. De plus, l'annotation automatique de ce gène ne prédit aucune séquence 5'UTR ni aucun séquence 3'UTR. L'hypothèse d'une mauvaise annotation du gène ACYPI080956 n'est donc pas à exclure. Ce cluster comporterait donc un seul gène de microARN exonique, api-mir-3043-3 et deux gènes de microARN introniques api-mir-3043-1 et api-mir-3043-2, tous les trois au sein des gènes hôtes ACYPI48456 et ACYPI082069.

Le troisième cluster (intronique et intergénique, Figure 2.11) inclut quatre gènes

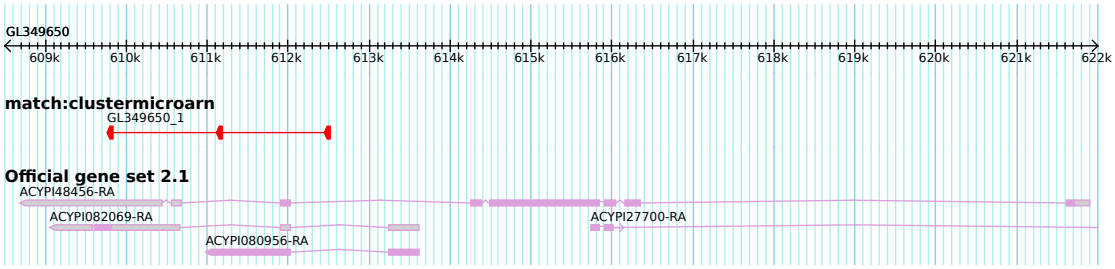


FIG. 2.10 – Région génomique contenant le cluster GL349650_1 des microARN api-mir-3043-1, api-mir-3043-2 et api-mir-3043-3 et les gènes hôtes ACYPI48456, ACYPI082069 et ACYPI080956. Le cluster de microARN est représenté en haut en rouge. Les flèches schématisent les gènes de microARN et le brin. Les gènes sont représentés en bas en violet. Les rectangles schématisent les exons (foncé pour les régions traduites, clair pour les régions non traduites), les traits les introns et la flèche le brin. La figure a été extraite d'AphidBase.

de microARN : api-mir-novel23-1, api-mir-novel23-2, api-mir-novel52b-2 et api-mir-novel52b-4 qui appartiennent à deux familles de microARN différentes avec la famille api-mir-novel23 qui est représentée au complet et la famille api-mir-novel52 pour laquelle il manque trois gènes : deux sur le même scaffold mais à une distance trop élevée pour être assimilé au même cluster et un sur un autre scaffold. De plus, les séquences entre ces deux familles sont très proches. Sur ces 4 gènes de microARN, 3 sont introniques au gène hôte ACYPI84584 (api-mir-novel52b-2, api-mir-novel23-2 et api-mir-novel23-1) et le dernier est intergénique (api-mir-novel52b-4). De plus, comme pour le deuxième cluster, les 4 gènes de microARN sont à des distances quasiment identiques : 1341 nucléotides, 1360 nucléotides et 1343 nucléotides. Le fait que les 4 gènes proviennent de deux familles très proches en séquence laisse supposer que ce cluster est annoté de façon correcte.

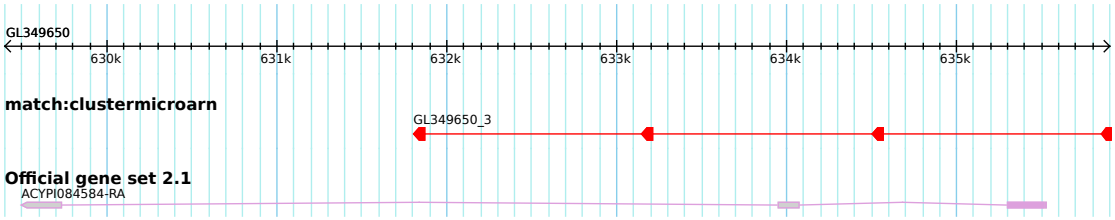


FIG. 2.11 – Région génomique contenant le cluster GL349650_3 des microARN api-mir-novel23-1, api-mir-novel23-2, api-mir-novel52b-2 et api-mir-novel52b-4 et le gène hôte ACYPI84584. Le cluster de microARN est représenté en haut en rouge. Les flèches schématisent les gènes de microARN et le brin. Les gènes sont représentés en bas en violet. Les rectangles schématisent les exons (foncé pour les régions traduites, clair pour les régions non traduites), les traits les introns et la flèche le brin. La figure a été extraite d'AphidBase.

En regardant de façon plus proche trois des clusters annotés avec différentes positions, il apparaît que les gènes de microARN qui constituent chacun des clusters

semblent provenir de la duplication d'un même gène ancestral ce qui laisse supposer que ces clusters sont annotés de façon correcte.

Duplication des gènes de microARN

Duplication Le génome du puceron du pois possède un nombre élevé de gènes d'ARNm dupliqués et un grand nombre de familles de gènes d'ARNm (environ 2.000) a subi une expansion [15]. Le nombre de séquences uniques de précurseurs de microARN prédits chez *Acyrthosiphon pisum* (329) inférieur au nombre de gènes (401) indique que, pour les microARN, il y a aussi eu une potentielle duplication ou expansion à des positions génomiques différentes de ces séquences de précurseurs.

Sur l'ensemble des 329 séquences uniques de précurseurs, 44 sont dupliquées avec pour un précurseur donné, un nombre de gènes dupliqués allant de 2 à 5 (voir Tableau 2.8).

nombre de copies sur le génome	nombre de précurseurs	nombre de gènes
1	285	285
2	27	54
3	10	30
4	3	12
5	4	20
total	329	401

Tableau 2.8 – Nombre de précurseurs uniques et nombre de gènes de microARN présents en 1, 2, 3, 4 ou 5 copies sur le génome.

L'ensemble des 116 gènes présents en plusieurs copies sur le génome représente 30 % des 401 gènes de microARN identifiés chez *A. pisum*. Parmi l'ensemble des clusters de gènes de microARN, seulement 3 sont constitués d'une même séquence répétée, les clusters incluant api-mir-3043-1/-2/-3, api-mir-novel52b-1/-3 et api-mir-novel46-2/-3. Les deux premiers clusters sont présents sur le même scaffold. Ce nombre représente environ 5 % du nombre total de clusters identifiés chez le puceron du pois, ce qui est en accord avec les observations faites chez d'autres espèces [116].

Familles de gènes de microARN au sein des clusters Les clusters de microARN sont constitués de gènes qui sont proches sur le génome et ces clusters sont souvent formés par des duplications en tandem des microARN [25]. De plus, des gènes de microARN qui sont issus d'une duplication ont de fortes chances de faire partie de la même famille car ils ont la même séquence d'origine. On peut donc se demander si les clusters identifiés précédemment sont principalement constitués d'une même famille. Le Tableau 2.9 montre le nombre de familles qui sont présentes dans les 52 clusters où sont présentes les familles avec qu'un seul gène. On peut voir que la majorité des clusters (46 %) inclue deux familles. Viennent ensuite les clusters incluant une, trois, quatre ou cinq familles et les clusters incluant six ou huit familles sont très peu représentés avec respectivement deux et un clusters. Les mêmes résultats sont présentés sur le Tableau 2.9 pour *D. melanogaster*. On peut voir que, en valeur absolue, les effectifs sont différents ce qui s'explique par le fait que le nombre total de clusters est différent, 52 pour *A. pisum* et 22

pour *D. melanogaster*. Néanmoins, les pourcentages et les répartitions sont globalement similaires entre les deux espèces.

nombre de familles présentes	nombre de clusters (%)	
	A. pisum	*D. melanogaster*
1	6 (12 %)	2 (9 %)
2	24 (46 %)	12 (54 %)
3	8 (15 %)	4 (18 %)
4	5 (9 %)	0 (0 %)
5	6 (12 %)	1 (5 %)
6	2 (4 %)	2 (9 %)
7	0 (0 %)	1 (5 %)
8	1 (2 %)	0 (0 %)
total	52 (100 %)	22 (100 %)

Tableau 2.9 – Tableau du nombre de familles présentes dans les clusters pour *A. pisum* et *D. melanogaster*.

Co-duplication des gènes de microARN et de leurs gènes hôtes Un grand nombre de gènes d'ARNm du puceron du pois sont dupliqués [15] et de plus il a été montré précédemment que 37 % (148) des gènes de microARN sont intragéniques (partie 2.2.2) et que au moins 30 % des gènes de microARN sont dupliqués (voir ci-dessus). La répartition entre les positions intergéniques, introniques et exoniques des 116 gènes issus de la duplication des 44 séquences de précurseurs est donnée ci-dessous :

- 79 gènes de microARN sont intergéniques ;
- 27 gènes de microARN sont introniques ;
- 10 gènes de microARN sont exoniques.

Il est possible que les 37 gènes de microARN intragéniques (27 introniques et 10 exoniques) et leurs gènes hôtes aient été co-dupliqués. Pour tester cette hypothèse, la base de données PhylomeDB [109] (version 4) a été utilisée. Sur l'ensemble des 22 gènes codant pour des protéines et incluant des gènes de microARN issus de duplication, seuls 6 d'entre eux (27 %) possèdent un arbre phylogénétique. Parmi ces 6 gènes hôtes, les deux gènes de microARN introniques api-mir-novel25-1 et api-mir-novel25-2 possèdent la même séquence et sont respectivement inclus dans les gènes hôtes ACYPI088736 et ACYPI007348. Ces deux gènes hôtes sont annotés comme issus d'une duplication sur PhylomeDB. La co-duplication d'un gène de microARN ancestral et de son gène hôte ancestral explique potentiellement l'apparition de gènes de microARN avec la même séquence au sein de gènes hôtes issus d'une même duplication. Le cas des gènes de microARN api-mir-novel25-1 et api-mir-novel25-2 co-dupliqués avec leurs gènes hôtes est le seul cas trouvé, à partir des données disponibles.

Il est possible que la duplication des gènes hôtes soit antérieure à la duplication des gènes de microARN, auquel cas l'apparition d'un gène de microARN se serait fait au sein de son gène hôte après la duplication de ce gène hôte. Une autre hypothèse serait que le gène de microARN se soit bien co-dupliqué avec son gène hôte mais que l'un des deux gènes de microARN possède maintenant une séquence différente dû à l'évolution

rapide des microARN. Néanmoins, les informations concernant la duplication des gènes étant fragmentaires, il se peut que certains gènes de microARN soient co-dupliqués avec leurs hôtes respectifs mais que cette information ne soit pas disponible.

2.2.3 Les interactions prédites entre microARN et ARNm

À partir des catalogues des ARNm et des microARN du puceron du pois, il est possible de prédire et proposer un catalogue des interactions entre ces molécules, en se basant sur les caractéristiques de ces interactions décrites dans la littérature.

Définitions

Nous distinguerons les graines, sites de fixation et interactions entre microARN matures et ARNm comme suit :

- *Graine* : la séquence en 5' du microARN qui est complémentaire d'une séquence sur le 3' UTR d'un transcrit. Celle-ci est nécessaire pour la fixation du microARN mature sur l'ARNm qui permet la répression de la traduction ou le clivage du transcrit de l'ARNm induit par le microARN. Cette séquence varie en taille et en position suivant les définitions, mais débute classiquement au 1er ou 2ème nucléotide du microARN et fait entre 6 et 7 nucléotides ;
- *Site de fixation* d'un microARN : la séquence sur le 3'UTR du transcrit de l'ARNm qui est complémentaire à la graine du microARN ;
- *Interaction* ou *couple* entre un microARN et un ARNm : une interaction entre un microARN mature et un ARNm (*interaction microARN/ARNm*) est prédite s'il existe au moins un site de fixation entre ce microARN mature et ce transcrit d'ARNm.

Pour une interaction microARN/ARNm, il peut y a avoir plusieurs sites de fixation de ce microARN mature sur l'ARNm.

À noter qu'il a été décidé de garder les distinctions de type « -1, -2 » obtenues sur les gènes entre les microARN matures, même s'ils sont identiques. C'est-à-dire que si api-mir-3051-1-5p cible ACYPI081796, alors api-mir-3051-2-5p ciblera aussi ACYPI081796.

Prédictions

Afin d'identifier les régulations post-transcriptionnelles des transcrits d'ARNm par les microARN matures, l'ensemble des sites de fixation des microARN matures sur les 3'UTR des transcrits d'ARNm doit être déterminé.

Il existe plusieurs méthodes de prédiction de ces sites de fixation. Afin de diminuer le nombre de fausses prédictions, il a d'abord été testé de prendre l'intersection de trois de ces méthodes : TargetScan (version 5, v5) [51], miRanda [57] et PITA [52]. Les seuils pour chacune des méthodes étaient choisis en prenant les seuils qui permettaient d'obtenir le ratio le plus élevé entre l'intersection des prédictions et le nombre de prédiction (l'indice de Jaccard [120]). Néanmoins, Ritchie *et al.* [121] ont montré que prendre l'intersection de plusieurs méthodes de prédictions ne permettait pas forcement d'avoir de meilleurs résultats. Il a donc été décidé de ne prendre que TargetScan v5 car c'est l'une des méthodes les plus strictes disponibles [47].

L'ensemble des sites de fixation a donc été prédit à partir des séquences des 3'UTR des gènes codant pour des protéines et des séquences des microARN matures en utilisant TargetScan v5. La description de cette méthode et les fichiers utilisés pour la prédiction ont été cités en 2.1.3. Les statistiques sur les prédictions du réseau d'interactions microARN/ARNm sont indiquées Tableau 2.10.

seuil	sites	microARN	ARNm	sites de fixation	couples
aucun	1	802 (100 %)	33.809 (100 %)	4.139.757 (100 %)	2.736.005 (100 %)
-0.3	1	802 (100 %)	31.964 (94,5 %)	1.162.561 (28,1 %)	961.915 (35,2 %)
aucun	2	801 (99,9 %)	26.382 (78 %)	2.125.692 (51,3 %)	740.280 (27,1 %)
-0.3	2	701 (87,4 %)	19.621 (58 %)	344.748 (8,3 %)	144.382 (5,3 %)

Tableau 2.10 – Résultat des prédictions de TargetScan avec ou sans seuil (seuil), avec 1 ou 2 sites de fixation minimum (sites). Sont indiqués le nombre de microARN matures (microARN) et d'ARNm concernés, le nombre de sites de fixation et le nombre de couples microARN/ARNm (couples). Est indiqué entre parenthèses le pourcentage d'éléments présents en fonction du nombre total d'éléments présents dans le réseau sans seuil.

Trois types de contraintes ont été appliqués alternativement pour obtenir les réseaux Tableau 2.10 :

1. Les sites de fixation doivent avoir un score inférieur ou égal à un seuil, ici -0.3 qui a été utilisé dans d'autres publications [122, 123] ;
2. Les couples microARN/ARNm doivent avoir un nombre de sites de fixation du microARN mature sur sa cible supérieur ou égal à un seuil, ici 2 sites de fixation ;
3. Les deux contraintes précédentes doivent être vérifiées.

La réduction du réseau aux sites de fixation avec un score inférieur ou égal à -0.3 permet de ne garder que les sites de fixation qui sont prédits comme induisant la plus forte répression de la traduction de l'ARNm et/ou réduction de sa demi-vie [51] et donc les sites de fixation qui ont potentiellement le plus d'influence. La filtration du réseau aux couples microARN/ARNm ayant au minimum 2 sites de fixation permet là aussi de garder des couples où la répression sera la plus forte dû aux effets possibles de coopération entre sites de fixation [51, 42] (voir l'Introduction pour plus de détails). De plus, du fait du nombre élevé de faux positifs (voir l'Introduction pour plus de détails) parmi les prédictions de sites de fixation, s'assurer qu'au moins 2 sites de fixation sont prédits entre un microARN mature et un ARNm permet de diminuer la probabilité que cette interaction soit fausse. L'utilisation conjointe des deux réductions permet de centrer encore plus le réseau sur les interactions qui ont le plus d'influence et de limiter le pourcentage de fausses interactions.

Pour le réseau sans restriction, tous les microARN matures prédits ciblent au minimum 1 ARNm et sont donc présents dans le réseau. De plus, 33.809 ARNm parmi les 35.462 (95 %) utilisés pour la prédiction des sites de fixations sont ciblés par au moins 1 microARN mature pour un nombre total de sites de fixation de 4.139.757 qui représente 2.736.005 couples.

Le premier seuil appliqué est la restriction sur le score des sites de fixation (2ème ligne du Tableau 2.10). Le réseau limité aux sites ayant un score global inférieur ou égal à -0.3 réduit grandement la combinatoire entre microARN matures et ARNm pour un nombre inchangé de microARN matures et un nombre élevé d'ARNm. En effet, les nombre des sites de fixation et de couples passent respectivement à 1.162.561 (28,1 %) et 961.915 (35,2 %) alors que le nombre de microARN et d'ARNm sont respectivement de 802 (100 %) et 31.964 (94,5 %).

Pour le réseau limité aux couples possédant 2 sites de fixations (3ème ligne du Tableau 2.10), le nombre de microARN et d'ARNm ne baisse que très faiblement : respectivement 801 (99,9 %) et 26.382 (78 %). Le nombre de couples, 740.280 (27,1 %), approche le nombre obtenu par la première réduction mais par contre le nombre de sites de fixation reste élevé, 2.125.692 (51,3 %). Ce qui est cohérent car l'on prend en compte uniquement les interactions microARN/ARNm avec deux sites de fixation

La troisième réduction est l'application des deux seuils (4ème ligne du Tableau 2.10, 2 sites de fixation avec pour les 2 un score global supérieur à -0.3). Cette réduction a l'effet le plus drastique en limitant grandement le nombre d'ARNm (19.621, 58 %), de sites de fixation (344.748, 8,3 %) et de couples (144.382, 5,3 %) et moins fortement le nombre de microARN matures (701, 87,4 %).

Sur l'ensemble des réductions appliquées au réseau, le nombre de microARN reste quasi constant. Le nombre d'ARNm varie surtout avec l'application des 2 seuils. Pour la combinatoire entre microARN matures et ARNm, les réductions ont toujours pour effet de diminuer fortement à la fois le nombre de sites de fixation et le nombre de couples.

Afin de comparer ces prédictions avec celles d'autres espèces, nous avons prédit par TargetScan v5 les sites de fixation non conservés des microARN matures chez *Drosophila melanogaster*. Les 3'UTR de *D. melanogaster* ont été téléchargés à partir de FlyBase [110] pour la dernière version du génome (version 6.01[3]), et les microARN à partir de miRBase. Les réseaux pour *A. pisum* et *D. melanogaster* ont été comparés avec un seuil de -0,3 sur le score. Le réseau de *D. melanogaster* contient 426 microARN matures, 26.547 ARNm, 237.023 sites de fixation et 216.264 couples. Sur ce réseau, le nombre moyen d'interactions est de 508 par microARN matures et de 8 par ARNm. Sur le réseau puceron du pois, il y a en moyenne 1.191 interactions par microARN matures et 30 par ARNm. On peut voir que les nombres d'interactions moyens sont supérieurs chez *A. pisum*.

Nous pouvons faire plusieurs hypothèses pour expliquer ces différences. On peut supposer que si un gène d'ARNm ancestral a été dupliqué et qu'il était ciblé par un microARN mature, alors les ARNm issus de sa duplication seront eux aussi ciblés par ce microARN mature. Et comme le nombre de duplications est important chez *A. pisum*, ceci pourrait en partie expliquer que les microARN matures du pucerons du pois ciblent quasiment deux fois plus en moyenne d'ARNm. Pour expliquer la différence du nombre moyen d'interactions par ARNm, rappelons que sur miRBase (pour drosophile) aucune distinction n'est faite entre deux séquences de microARN matures identiques alors qu'ici pour le puceron du pois, nous avons gardé la distinction de type « -1, -2 » pour les microARN matures. Ceci implique que dans le cas de deux microARN matures identiques, deux interactions sont notées dans le réseau chez *A. pisum*, et une seule pour *D. melanogaster*. De plus, les niveaux et qualités d'annotations diffèrent grande-

[3] `ftp://ftp.flybase.net/genomes/Drosophila_melanogaster/current/fasta/`

ment entre la drosophile et le puceron du pois. La présence de microARN ou de gènes mal annotés pourrait aussi expliquer ces différences. Dans tous les cas, le nombre plus élevé d'interactions microARN/ARNm chez *A. pisum* ouvre un champ important de modification de l'expression génique chez cette espèce.

2.3 Résumé et conclusion : l'identification et l'analyse d'un nouveau catalogue de microARN chez *A. pisum*

Ce chapitre décrit le catalogue des microARN du pucerons du pois en utilisant les nouvelles banques de séquences obtenues sur des embryons synchronisés du puceron du pois à l'avenir asexués ou sexués. Sur l'ensemble des 7 espèces du phylum Hexapoda dont fait partie *Acyrthosiphon pisum*, le nombre de précurseurs de microARN et le nombre de microARN matures du puceron du pois font partie des plus élevés. Les 329 précurseurs identifiés correspondent à 401 gènes de microARN, ce qui indique une duplication des gènes de microARN. L'hypothèse d'un nombre important de duplications est soutenue par la présence de nombreuses familles de microARN, comparée à *Drosophila melanogaster*, et le fait qu'elles soient quasi uniquement constituées de gènes identifiés uniquement chez *A. pisum*. Cette duplication et la divergence entre gènes de microARN après duplication pourraient être à l'origine du nombre plus élevé de précurseurs uniques de microARN et microARN matures uniques chez le puceron du pois.

Plus de la moitié des gènes de microARN du puceron du pois (51,6 %) est réparti en clusters génomiques. Cette répartition en clusters des gènes de microARN chez le puceron est différente de celle observée chez *Drosophila melanogaster*. En effet, le pourcentage de clusters de « petite taille », entre 2 et 3 gènes, est plus important chez *D. melanogaster*, alors que chez *A. pisum* les clusters de taille « moyenne », entre 4 et 6 gènes, sont plus abondants. Il est aussi apparu que le puceron possède trois clusters de « grande taille » avec 12, 14 et 16 gènes de microARN. L'alignement montre que pour ces trois clusters, les microARN appartiennent principalement aux mêmes familles qui ont des séquences quasiment identiques. Ces clusters découleraient donc de la duplication d'un ou plusieurs gènes ancestraux, ce qui expliquerait le plus petit pourcentage de clusters de « petite taille » car le taux de duplication plus élevé induirait des clusters plus grands en nombre de gènes. De plus, on a pu voir que la répartition en familles de gènes de microARN au sein des clusters étaient quasiment identique entre *A. pisum* et *D. melanogaster*.

Ce chapitre introduit aussi le premier réseau d'interaction entre les microARN matures et les ARNm du puceron du pois précédemment identifiés [15, 78] et mis à jour sur la version 2 du génome. Sans restriction, ce réseau prédit à l'aide de TargetScan [51] implique l'ensemble des 802 microARN matures (issus des 401 gènes de microARN) et 33.809 ARNm (ce qui représente 95 % des ARNm initialement utilisés pour la prédiction) pour un nombre de sites de fixation de 4.139.757 et un nombre de couples de 2.736.005. Les différentes réductions opérées sur ce réseau permettent de limiter le nombre de sites de fixation et le nombre d'interactions respectivement à 7 % et 4 % du nombre initial pour l'association des 2 filtres tout en gardant un nombre élevé de microARN et d'ARNm respectivement à 83 % et 47% du nombre initial pour l'association des 2 filtres.

Le réseau d'interactions microARN/ARNm obtenu, avec ou sans restriction, implique l'ensemble des microARN matures identifiés dans des embryons d'*Acyrthosiphon pisum* et l'ensemble des ARNm identifiés dans la version 2 du génome. Afin de mettre en évidence les différences dans les processus de régulation des ARNm par les microARN entre le développement d'embryons à l'avenir asexué et le développement d'embryons à

l'avenir sexué, le réseau doit être réduit aux microARN matures et ARNm dont l'expression évolue de façon différente au cours du développement des deux morphes étudiés. La méthode utilisée pour extraire les microARN matures et les ARNm à ceux possédant des cinétiques différentes ainsi que leurs analyses et le réseau qui découle de cette réduction sont présentées dans le chapitre suivant.

Chapitre 3

Analyse, classification et comparaison des cinétiques d'expression des ARNm et microARN au cours du développement des embryons sexués et asexués du puceron du pois

Le chapitre précédent présente un réseau d'interactions entre les ARNm et les microARN matures du puceron du pois. Sur l'ensemble de ce réseau, seulement certains éléments sont différentiellement régulés au cours du temps entre l'embryogenèse sexuée et l'embryogenèse asexuée. Ce chapitre présente la méthode utilisée pour identifier et analyser les ARNm et les microARN différentiellement régulés entres les deux embryogenèses ainsi que la réduction du réseau d'interaction à ces éléments.

3.1 Matériels et méthodes : discrétisation des cinétiques et classification des transitions cinétiques

Cette partie présente la façon dont les cinétiques d'expression des ARNm et microARN ont été analysées, discrétisées et comparées.

3.1.1 Description du paquet R `edgeR`

`edgeR` [124, 125, 126, 127] est une suite logicielle développée pour des analyses statistiques R [128]. Elle permet, entre autres, de faire une analyse différentielle entre deux conditions pour des données issues de séquençage haut-débit (transcrits, exons, etc) que nous appellerons *éléments* dans la suite pour rester générique. `edgeR` modélise par un loi binomiale négative l'expression des éléments identifiés par le séquençage. Cette binomiale négative possède deux paramètres : la valeur moyenne d'expression, qui correspond à l'abondance de l'élément dans l'échantillon, et la variance, dont la valeur dépend de la moyenne et d'un valeur de dispersion (voir plus loin). Dans notre cas, l'analyse différentielle à l'aide de `edgeR` se fait en quatre étapes :

1. Calcul des facteurs de normalisation des tailles brutes des banques de séquençage par rapport à une banque de référence en utilisant la méthode « trimmed mean of M-values » (TMM) développée par Robinson et Oshlack [129]. Cette méthode part du principe que peu d'éléments sont différentiellement exprimés. Elle calcule donc les facteurs qui permettent de minimiser le log de l'accroissement (log-fold change) pour la majorité des éléments sur l'ensemble des banques en supprimant les éléments fortement exprimés ou avec un rapport logarithmique de probabilité (log-ratio) élevé dans l'analyse. La fonction utilisée est : `calcNormFactors` ;
2. Estimation de la dispersion entre les différentes banques avec une valeur de dispersion identique pour l'ensemble des éléments dans les banques. C'est-à-dire que la différence entre le paramètre de variance pour chacun des éléments ne dépendra que de la valeur de la moyenne. La fonction utilisée est : `estimateCommonDisp` ;
3. Estimation de la dispersion spécifique à chacun des éléments. Afin de pallier le nombre souvent faible d'échantillons, la dispersion de chaque élément est « tirée » vers la dispersion commune. La fonction utilisée est : `estimateTagwiseDisp` ;
4. Pour chacun des éléments, un test de différence de moyenne entre deux groupes de variables dont les distributions suivent une loi binomiale négative est effectué. Ce test développé par Robinson et Smyth [125] permet d'associer une p-value à chacune des comparaisons entre éléments pour deux groupes, c'est-à-dire deux conditions biologiques dans notre cas. La fonction utilisée est : `exactTest`.

L'ensemble de ces fonctions a été utilisé avec les paramètres par défaut.

3.1.2 Discrétisation des cinétiques d'expression

Afin de discrétiser les cinétiques sexuées et asexuées, le test statistique que nous venons de décrire a été appliqué entre les expressions aux temps T_i et T_{i+1} à la fois pour les ARNm d'une part, et les microARN d'autre part, prenant en compte les répétitions biologiques et techniques (voir chapitre 2.1.2 tableau 2.2). Pour chaque paire de pas de temps consécutifs, un test statistique de différence de moyennes entre les deux

échantillons suivant une loi binomiale négative a été effectué par `edgeR`. L'ensemble des p-values a été ajusté par la méthode de Benjamini et Hochberg [130]. L'expression entre les temps T_i et T_{i+1} a été considérée comme différente quand la p-value ajustée associée au test entre ces deux temps était inférieure ou égale au seuil de 5 %. Une fois les différences significatives entre l'ensemble des temps T_i et T_{i+1} obtenues, ces différences ont été discrétisées de la façon suivante :

- 0 si la différence d'expression entre T_i et T_{i+1} n'est pas significative ;
- 1 si la différence d'expression entre T_i et T_{i+1} est significative et que l'expression croît en fonction du temps ;
- -1 si la différence d'expression entre T_i et T_{i+1} est significative et que l'expression décroît en fonction du temps.

Pour les ARNm, la discrétisation a été effectuée sur le comptage des 36.990 séquences d'ARNm pour l'ensemble des 19 banques de séquençage (les comptages pour les trois réplicats de séquençage pour le temps T_0 ont été sommés). La discrétisation pour les cinétiques des microARN a été effectuée sur le comptage des 573 séquences uniques de microARN matures pour l'ensemble des 21 banques de séquençage (les comptages pour les trois réplicats de séquençage pour chacun des réplicats biologiques ont été sommés). Pour les ARNm et les microARN matures, le même protocole a été appliqué. Les résultats sur ces discrétisations sont résumés dans les Tableaux 3.2 et 3.5 dans la partie 3.2.1. Ils présentent les valeurs discrètes associées respectivement aux ARNm et aux microARN matures entre chaque pas de temps consécutif, pour les profils sexués et asexués séparément.

Pour chacun des ARNm et des microARN matures, la dynamique d'expression est représentée par un couple de profils de trois valeurs discrètes prenant leur valeur dans {-1,0,1}, l'un pour la cinétique sexuée et l'autre pour la cinétique asexuée. Par exemple le profil sexué S={0,1,-1} d'un microARN ou d'un ARNm avec un seuil de 5 % signifie que les expressions aux temps T_0 et T_{1S} ne sont pas significativement différentes, que l'expression au temps T_{1S} est statistiquement plus faible que l'expression au temps T_{2S} et que l'expression au temps T_{2S} est statistiquement plus élevée que l'expression au temps T_{3S}. Dans la suite de la thèse, le mot « profil » sera utilisé pour faire référence à ces profils sexués (S) ou asexués (A) associés aux ARNm et microARN matures.

3.1.3 Classification des transitions cinétiques

Afin de caractériser les transitions des cinétiques sexuées vers les cinétiques asexuées pour chacun des ARNm et des microARN matures, un ensemble de neuf règles qui généralisent et classifient les différences entre le caractère sexué et le caractère asexué a été défini. Ces règles ont pour but d'aider à l'interprétation, c'est-à-dire de distinguer des cinétiques comparables du point de vue de la régulation. Elles permettent trois choses :

1. Partitionnement des couples de profils ;
2. Prise en compte d'une évolution monotone, transitoire ou encore d'un décalage temporel ;
3. Regroupement possible des règles par paires *positives*/*négatives*.

Ces règles caractérisent le changement de la cinétique sexuée vers la cinétique asexuée, dans cet ordre. Elles ont été définies pour couvrir l'ensemble des $3^6 - 3^3 = 702$

combinaisons possibles pour les six valeurs associées à un élément et où il y a au moins une différence entre le profil S et A pour une $i^{\text{ème}}$ valeur : $\exists i \in \{1,2,3\}\ v_i^s \neq v_i^a$. La partie qui suit décrit ces neuf règles, les classe positivement ou négativement et donne entre parenthèse le nombre de combinaison de profils qui sont couverts par chacune des règles. On note $regle(S,A)$ si le couple de profil S,A correspond à la règle $regle$.

Deux règles sont définies pour une évolution monotone, une positive : « augmente » (189) et une négative : « diminution » (189) :

$$\begin{aligned} augmente(S,A) &\Leftrightarrow \forall v_i^s, v_i^a, i \in \{1,2,3\}\ v_i^s \leq v_i^a; \\ diminution(S,A) &\Leftrightarrow \forall v_i^s, v_i^a, i \in \{1,2,3\}\ v_i^s \geq v_i^a. \end{aligned}$$

Deux règles sont définies pour un décalage temporel, une positive : « avance » (18) et une négative : « retard » (18), où $x_1, x_2 \in \{-1,1\}$:

$$\begin{aligned} avance(S,A) \Leftrightarrow S &= (0,0,x_1) \wedge A = ((0,x_1,0) \vee (x_1,0,0)) \vee \\ S &= (0,x_1,0) \wedge A = (x_1,0,0) \vee \\ S &= (0,x_1,x_2) \wedge A = ((x_1,0,x_2) \vee (x_1,x_2,0)) \vee \\ S &= (x_1,0,x_2) \wedge A = (x_1,x_2,0). \\ retard(S,A) \Leftrightarrow S &= (0,x_1,0) \wedge A = (0,0,x_1) \vee \\ S &= (x_1,0,0) \wedge A = ((0,x_1,0) \vee (0,0,x_1)) \vee \\ S &= (x_1,0,x_2) \wedge A = (0,x_1,x_2) \vee \\ S &= (x_1,x_2,0) \wedge A = ((0,x_1,x_2) \vee (x_1,0,x_2)). \end{aligned}$$

Quatre règles ont été définies pour une évolution transitoire, deux positives : « apparition pic positif » (7) et « disparition pic négatif » (7) et deux négatives : « apparition pic négatif » (7) et « disparition pic positif » (7), où $x \in \{-1,0,1\}$:

$$\begin{aligned} apparitionPicPositif(S,A) \Leftrightarrow S &= (x,0,0) \wedge A = (x,1,-1) \vee \\ S &= (0,0,x) \wedge A = (1,-1,x) \vee \\ S &= (0,0,0) \wedge A = (1,0,-1). \\ disparitionPicNegatif(S,A) \Leftrightarrow S &= (x,-1,1) \wedge A = (x,0,0) \vee \\ S &= (-1,1,x) \wedge A = (0,0,x) \vee \\ S &= (-1,0,1) \wedge A = (0,0,0). \\ apparitionPicNegatif(S,A) \Leftrightarrow S &= (x,0,0) \wedge A = (x,-1,1) \vee \\ S &= (0,0,x) \wedge A = (-1,1,x) \vee \\ S &= (0,0,0) \wedge A = (-1,0,1). \\ disparitionPicPositif(S,A) \Leftrightarrow S &= (x,1,-1) \wedge A = (x,0,0) \vee \\ S &= (1,-1,x) \wedge A = (0,0,x) \vee \\ S &= (1,0,-1) \wedge A = (0,0,0). \end{aligned}$$

Une règle, « défaut » (260), est associée par défaut à tous les couples de profils qui ne correspondent à aucune des huit règles précédentes. Le Tableau 3.1 redonne une définition des règles et la Figure 3.1 illustre ces règles.

Ces règles sont mutuellement exclusives, c'est-à-dire qu'un couple de profils ne sera associé qu'à une et une seule règle afin d'aider à l'interprétation. Pour chacun des ARNm et des microARN matures, la règle à laquelle correspond le couple de profils sera associée à cet ARNm ou ce microARN mature.

règle	description (sexué vers asexué)	exemple	
		sexué	asexué
augmentation	Au moins l'une des trois valeurs augmente et les autres sont identiques.	-1, 0, -1	-1, 1, 0
diminution	Au moins l'une des trois valeurs diminue et les autres sont identiques.	-1, 0, -1	-1, -1, -1
avance	Une ou deux valeurs sont avancées. Ce/ces valeurs remplacent un zéro ou une des valeurs qui bougent.	0, 1, 1	1, 1, 0
retard	Une ou deux valeurs sont retardées. Ce/ces valeurs remplacent un zéro ou une des valeurs qui bougent.	-1, -1, 0	-1, 0, -1
apparition pic positif	Un pic d'expression positif apparaît.	-1, 0, 0	-1, 1, -1
apparition pic négatif	Un pic d'expression négatif apparaît.	-1, 0, 0	-1, -1, 1
disparition pic positif	Un pic d'expression positif disparaît.	1, 0, -1	0, 0, 0
disparition pic négatif	Un pic d'expression négatif disparaît.	-1, 1, 1	0, 0, 1
défaut	Le couple ne correspond à aucune autre règle.	0, 1, 1	1, 0, 0

Tableau 3.1 – Tableau décrivant les règles permettant de partitionner en neuf classes les différences entre le profil sexué et le profil asexué pour les ARNm et les microARN matures. Pour chacune des règles, une description est donnée ainsi qu'un exemple.

3.1.4 Étude des ARNm différentiellement exprimés enrichis en annotations fonctionnelles

Obtention des ARNm différentiellement exprimés enrichis en annotations fonctionnelles par Blast2GO

Blast2GO [97, 98] permet, en plus des fonctionnalités décrites précédemment, de calculer un enrichissement en annotations fonctionnelles d'un sous-ensemble d'ARNm d'intérêt possédant une annotation fonctionnelle (ici les ARNm qui seront différentiellement exprimés) contre l'ensemble des ARNm annotés (la liste de référence). L'annotation utilise des termes de l'ontologie GO. Pour savoir si une annotation GO donnée est enrichie dans une liste d'ARNm spécifique, Blast2GO effectue un test exact de Fisher entre d'une part le ratio entre le nombre d'ARNm possédant cette annotation et ceux ne possédant pas cette annotation dans le sous-ensemble d'ARNm d'intérêt, contre d'autre

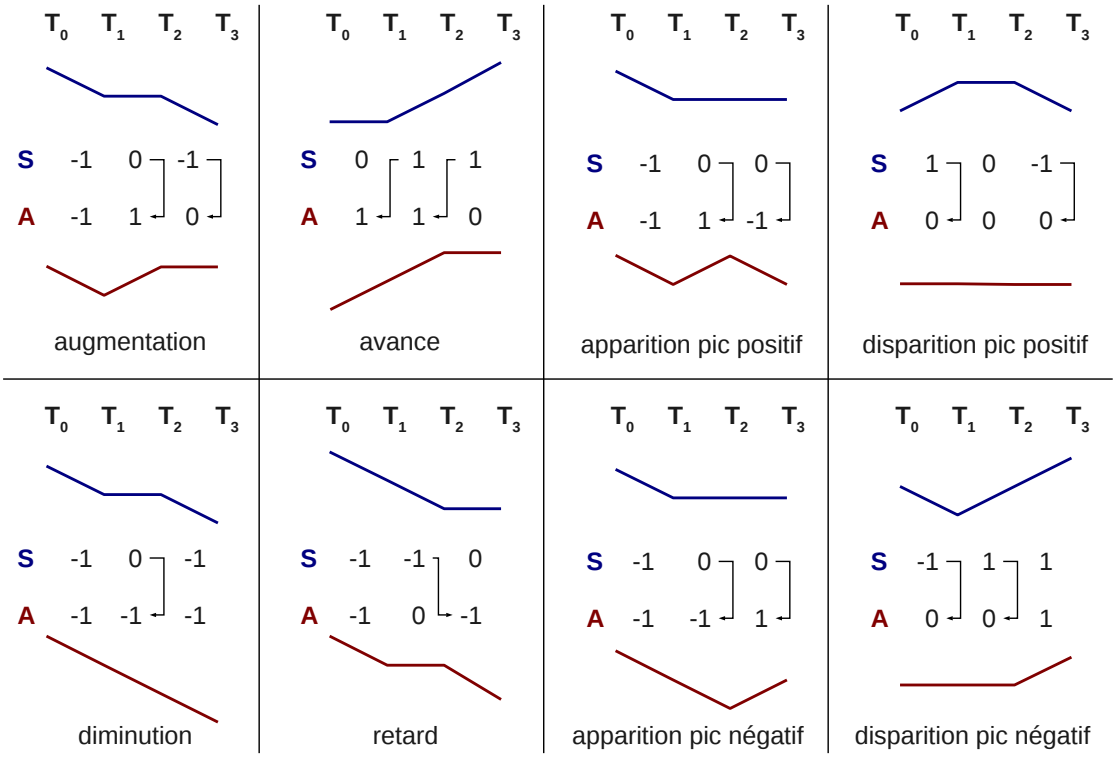


FIG. 3.1 – Schéma des couples de profils des cinétiques sexuées et asexuées associés aux principales règles de transitions décrites dans le Tableau 3.1. Pour chacun des huit schémas illustrant les huit règles, les quatre temps T_0, T_1, T_2, T_3 sont visibles en haut et le nom de la règle en bas. Le profil sexué est schématisé en bleu et par la lettre S et le profil asexué est schématisé en rouge et par la lettre A. Pour chacune des règles, des flèches lient les valeurs qui changent entre le profil sexué et asexué et qui caractérisent le classement du couple de profils par cette règle.

part ce même ratio mais obtenu sur l'ensemble des ARNm de référence. Les différents tests exacts de Fisher ont été réalisés ici en utilisant l'interface graphique de Blast2GO[1] avec un seuil sur les p-values ajustées à 5 %.

Visualisation et analyse des ARNm différentiellement exprimés enrichis en annotations fonctionnelles

Afin de supprimer la redondance, de résumer et de visualiser les résultats obtenus sur les annotations GO qui sont enrichies, le logiciel REViGO[2] [131] a été utilisé. REViGO fournit, à partir d'une liste de termes GO, un ensemble de clusters de termes GO similaires. Cela permet d'obtenir une partition des termes en fonction de leur similarité.

Étant donné un ensemble de termes GO enrichis avec une p-value associée, REViGO va pour chaque paire de termes calculer une similarité : celle par défaut et qui a été utilisée ici est la similarité SimRel [132]. Elle renvoie une valeur comprise entre 0 et 1,

[1] Disponible à cette adresse : `http://www.blast2go.com/b2ghome`
[2] `http://revigo.irb.hr/`

où une valeur de 1 signifie que les termes sont similaires. En partant des deux termes t_i et t_j les plus similaires, REViGO garde les deux termes s'ils ont un degré de similarité inférieur à un certain seuil (0,7 par défaut). Sinon un ensemble de règles est utilisé pour choisir quel terme est supprimé. Un résumé des différentes étapes et des règles de choix sont présenté Figure 3.2.

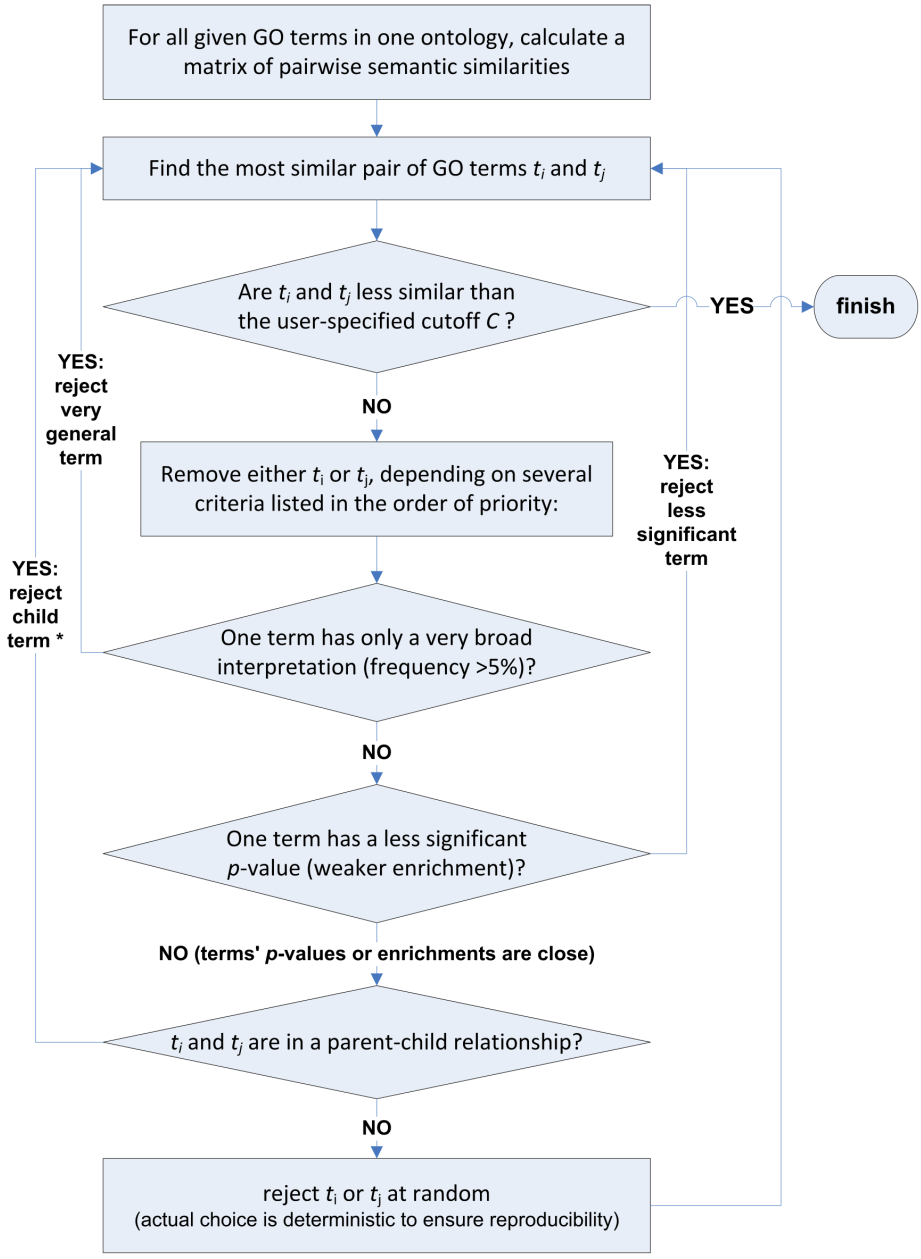


FIG. 3.2 – Les différentes étapes et règles de la méthode REViGO. *Si le terme fils représente 75 % ou plus de l'ensemble des termes fils u terme parent, alors le terme parent est supprimé à la place du terme fils. Figure tirée de Supek *et al.* [131].

Les termes qui subsistent à la fin de cette procédure sont les termes qui représente-

ront les clusters de termes GO et où un terme est associé à un cluster si sa similarité est la plus élevée pour le terme représentant ce cluster. REViGO permet de visualiser ces clusters notamment par une visualisation par un diagramme de dispersion. Ce diagramme est obtenu en utilisant la similarité entre les clusters et en les projetant sur deux axes, appelés axes sémantiques, à l'aide entre autre du multidimensional scaling (MDS).

3.2 Résultats : Comparaison des expressions géniques entre embryons sexués et asexués

L'objectif des expériences réalisées vise à comparer des profils d'expression des ARNm et des microARN matures dans des embryons du puceron du pois engagés dans un développement sexué ou asexué. Ceci permettra par la suite de réduire le graphe d'interactions aux éléments d'intérêt : on ne retient que les interactions entre des microARN et des ARNm possédant des cinétiques différentes entre les développements sexués et asexués. Le protocole expérimental présenté précédemment (voir Figure 2.1) indique la façon d'obtenir des cinétiques de développement embryonnaire synchrones entre les embryogenèses sexuées et asexuées. Les échantillons obtenus correspondent à différents stades embryonnaires (du Stade 17 ou Stade 20, voir Tableau 2.2) à la fois des embryons sexués et asexués.

La Figure 3.3 ci-dessous décrit les deux cinétiques d'embryogenèses ainsi obtenues. Le temps T_0 correspond au stade 17 qui est le stade dit flexible où l'application de kinoprène permet de diriger le devenir d'embryons à l'avenir sexué en embryons à l'avenir asexué. Les extractions aux stades 18, 19 et 20 sont les extractions obtenues à la fois chez les pucerons sans application de kinoprène (sexués : S) et avec application de kinoprène (asexués : A). Le temps T_0 est donc commun aux deux cinétiques et les temps T_1, T_2 et T_3 sont spécifiques aux cinétiques. Les temps S=$\{T_0,T_{1S},T_{2S},T_{3S}\}$ constituent la cinétique sexuée et les temps A=$\{T_0,T_{1S},T_{2S},T_{3S}\}$ constituent la cinétique asexuée.

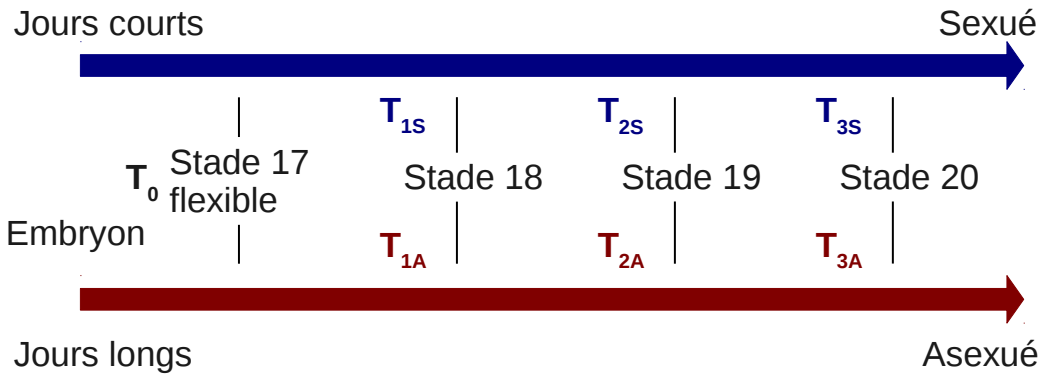


FIG. 3.3 – Schéma des cinétiques obtenues sur des embryogenèses de puceron femelles à l'avenir sexué ou asexué pour les stades 17, 18, 19 et 20. En bleu : l'embryogenèse sexuée (S) ; en rouge l'embryogenèse asexuée (A) ; les temps T_0, T_{1S}, T_{2S}, T_{3S} et T_{1A}, T_{2A}, T_{3A} représentent respectivement les extractions aux stades 17, 18 sexué, 19 sexué, 20 sexué et 18 asexué, 19 asexué, 20 asexué.

Cette partie présente tout d'abord l'identification des ARNm et des microARN qui sont différentiellement régulés entre les embryogenèses sexuées et asexuées. Par la suite, il est décrit comment la réduction de ces deux catalogues permet de limiter le nombre d'interactions du réseau.

3.2.1 Identification et classification des ARNm et des microARN matures régulés

Cette étude consiste à comparer l'expression des ARNm et des microARN pendant le développement de deux types d'embryons. Afin d'aider à l'interprétation, il a été décidé de discrétiser l'expression entre chaque paire de pas de temps consécutifs afin de rendre comparable les cinétiques d'expression.

Identification et classification des ARNm

Identification des ARNm différentiellement régulés Pour l'ensemble des ARNm et pour les deux cinétiques, sexuée : S={T_0,T_{1S},T_{2S},T_{3S} } et asexuée : A={T_0,T_{1S},T_{2S},T_{3S} }, une discrétisation de l'évolution de l'expression entre chaque paire de pas de temps consécutifs a été effectuée (voir partie 3.1.2). Les résultats obtenus pour l'ensemble des ARNm sont présentés dans le Tableau 3.2 ci-dessous.

	sexué (% de variation pour le profil asexué)		
valeur	T_0/T_{1S}	T_{1S}/T_{2S}	T_{2S}/T_{3S}
0	27.397 (-0,1%)	32.969 (+0,2%)	31.005 (+1,4%)
-1	5.585 (-0,2%)	1.190 (+3,4%)	2.625 (-15,6%)
1	4.008 (+1,2%)	2.931 (-7,7%)	3.360 (-0,7%)
total (-1/1)	9.593 (+0,4%)	4.121 (-4%)	5.985 (-7,2%)

Tableau 3.2 – Nombre d'ARNm associés aux valeurs discrètes -1, 0 et 1 pour l'ensemble des paires de pas de temps consécutifs des cinétiques sexuées. Pour les cinétiques asexuées, le pourcentage de différence est présenté entre parenthèses. L'analyse a été faite sur l'ensemble des 36.990 séquences uniques d'ARNm. La dernière ligne représente la somme des lignes -1 et 1.

On peut voir que la répartition des ARNm en -1, 0 et 1 pour des temps consécutifs identiques est semblable entre la cinétique sexuée et la cinétique asexuée. Le nombre d'ARNm ayant une valeur différente de zéro varie entre environ 10 % (différence entre $T_{1S/2A}$ et $T_{2S/2A}$) et 26 % ($T_0/T_{1S/1A}$). Cette différence de répartition nous indique que les plus grandes variations d'expression se font entre le temps T_0 et le temps T_1 pour la cinétique sexuée et la cinétique asexuée. Le temps T_0 est le dernier temps pendant lequel l'embryon est sensible au kinoprène, où son avenir n'est pas encore déterminé, ce qui veut dire que le moment où l'on observe le plus d'ARNm dont l'expression évolue est le moment où l'embryon passe d'un état flexible à un état où il ne l'est plus. De plus, les pourcentages de variation entre l'embryogenèse sexuée et asexuée sont globalement faibles. Sauf pour la transition T_{2S}/T_{3S} valeur -1 (baisse de l'expression) qui pourrait montrer une régulation différente au dernier stade de type répression supérieure dans le cas sexué.

Pour chacun des ARNm, deux profils cinétiques discrétisés sont obtenus : l'un pour le développement sexué et le second pour le développement asexué. Pour résumer : pour chaque ARNm, on définit ainsi à la fois un vecteur (profil) de trois valeurs pour la cinétique sexuée et un vecteur pour la cinétique asexuée. Ce sont ces deux profils qui sont ensuite comparés. Dans la suite de la thèse, le mot profil est utilisé pour faire référence à ces profils sexués ou asexués associés aux ARNm et composés de valeurs

discrètes. Parmi les 36.990 ARNm, 4.996 ARNm (13,5 %) ont un profil sexué qui est différent du profil asexué par au moins une valeur. Par la suite, nous décrirons ces ARNm comme ayant des cinétiques différentielles.

Enrichissement fonctionnel des ARNm différentiellement régulés Au sein des ARNm ayant des cinétiques différentielles, 1.640 possèdent une annotation GO. À l'aide du logiciel Blast2GO (voir partie 3.1.4), un enrichissement fonctionnel a été réalisé sur ces 1.640 ARNm contre la liste de référence comprenant 10.062 ARNm (nombre des ARNm du pucerons du pois ayant une annotation GO). Sur les 2.340 annotations GO différentes pour les ARNm ayant des cinétiques différentielles, 39 sont enrichies dont 25 dans la classe processus biologique (biological process, BP), deux dans la classe fonction moléculaire (molecular fonction, MF) et 12 dans la classe composante cellulaire (cellular component, CC). Au vu du faible nombre dans la classe MF et du caractère peu informatif des annotations de la catégorie CC dans cette étude, seule une analyse par REViGO des annotations BP a été faite, avec les paramètres par défaut. La Figure 3.4 représente l'ensemble des clusters obtenus sur BP avec les termes les représentant, et le Tableau 3.3 représente l'ensemble des annotations GO BP enrichies ainsi que les $\log_{10}$ p-value qui ont été obtenues par Blast2GO.

Les 25 annotations sont regroupées en neuf groupes, avec les cluster 1, 5, 7 et 9 qui regroupent plusieurs termes. Le cluster 1, incluant quasiment la moitié des termes, regroupe des termes associés à la régulation négative de l'expression des gènes, de la transcription et de certaines branches du métabolisme. Le second cluster en importance par rapport au nombre de termes, le cluster 5, regroupe des termes concernant le développement et notamment le développement du système nerveux. Le cluster 7, qui bénéficie d'un fort support en ARNm annoté (plus de 200), concerne des termes liés au de développement anatomique. Le cluster 9 concerne des termes liés à la transcription par l'ARN polymérase II.

On peut voir qu'un certain nombre de termes concernent soit la régulation de la transcription, soit le développement. Cette dernière fonction est cohérente vue l'expérimentation menée : la réduction des ARNm à ceux possédant des cinétiques différentes entre le développement d'embryons sexués et le développement d'embryons asexués. En nombre d'ARNm annotés, on observe une forte signature du cluster 7. De même pour le cluster 5 où seul le terme « central nervous system development » est faiblement représenté en nombre d'ARNm annotés. Pour le cluster 1, les termes sont globalement peu représentés mais de façon uniforme. Le cluster 9 possède la plus faible signature en terme de nombre d'ARNm.

Il semble que parmi les ARNm ayant des cinétiques différentielles, un nombre élevé soit annoté par des fonctions ayant un lien avec le développement et des fonctions de régulation.

Classification des ARNm différentiellement régulés Les 4.996 ARNm ayant des cinétiques différentielles ont ensuite été classés afin de comparer l'ensemble des profils des développements embryonnaires sexués et asexués. La méthode utilisée pour classer les ARNm avec des cinétiques différentielles à l'aide de règles définis *a priori* est décrite partie 3.1.3. Le Tableau 3.4 présente le nombre d'ARNm associé à ces règles.

On peut voir que les deux règles dominantes sont les règles « augmentation » et

cluster	GO ID	description	$\log_{10}$ p-value	nombre
1	GO :0009892	negative regulation of metabolic process	-2.4338	60
	GO :0031324	negative regulation of cellular metabolic process	-1.8670	47
	GO :0045892	negative regulation of transcription, DNA-dependent	-2.4869	41
	GO :0010605	negative regulation of macromolecule metabolic process	-2.2394	57
	GO :0051253	negative regulation of RNA metabolic process	-2.3582	41
	GO :0031327	negative regulation of cellular biosynthetic process	-2.0744	42
	GO :0010629	negative regulation of gene expression	-2.4026	50
	GO :0009890	negative regulation of biosynthetic process	-1.9400	42
	GO :0010558	negative regulation of macromolecule biosynthetic process	-2.0744	42
	GO :2000113	negative regulation of cellular macromolecule biosynthetic process	-2.1118	42
	GO :0051172	negative regulation of nitrogen compound metabolic process	-2.4869	43
	GO :0045934	negative regulation of nucleobase-containing compound metabolic process	-2.4869	43
2	GO :0032501	multicellular organismal process	-3.6108	308
3	GO :0032502	developmental process	-3.0293	288
4	GO :0044699	single-organism process	-3.0416	584
5	GO :0044707	single-multicellular organism process	-3.0171	290
	GO :0048731	system development	-1.8974	212
	GO :0007417	central nervous system development	-2.5578	47
	GO :0007275	multicellular organismal development	-2.8778	257
6	GO :0006366	transcription from RNA polymerase II promoter	-1.6489	70
7	GO :0044767	single-organism developmental process	-2.4869	216
	GO :0048856	anatomical structure development	-2.4869	255
8	GO :0044763	single-organism cellular process	-1.6489	498
9	GO :0006357	regulation of transcription from RNA polymerase II promoter	-2.1844	64
	GO :0000122	negative regulation of transcription from RNA polymerase II promoter	-2.0852	29

Tableau 3.3 – Synthèse des résultats obtenus par REViGO avec les paramètres par défaut sur la clusterisation des termes GO BP suivant leur similarité. Les valeurs de $\log_{10}$ p-value sont obtenues à partir des résultats de Blast2GO et la colonne « nombre » donne le nombre d'ARNm annotés avec le terme considéré. Le premier terme de chaque cluster représente le groupe.

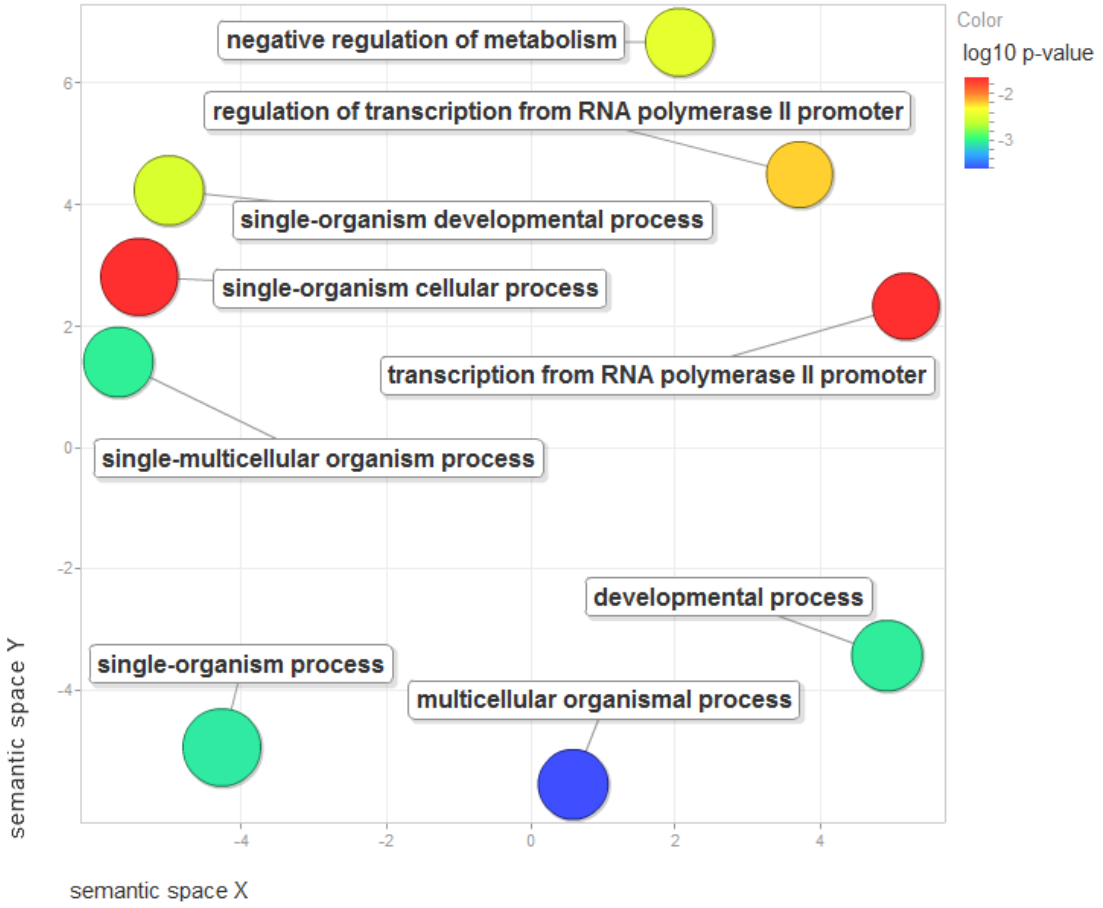


FIG. 3.4 – Figure des clusters de termes GO BP obtenue par REViGO avec les paramètres par défaut. Le nom affiché est celui des termes représentant les clusters et les valeurs de $\log_{10}$ p-value sont obtenues à partir des résultats de Blast2GO.

règle	nombre
augmentation	2434
diminution	2256
avance	**134**
retard	**75**
apparition pic négatif	25
disparition pic négatif	22
disparition pic positif	20
défaut	19
apparition pic positif	11
total	4.996

Tableau 3.4 – Distribution décroissante du nombre d'ARNm parmi les ARNm différentiellement régulés associés à chacune des règles décrites partie 3.1.3.

« diminution » qui regroupent à elles seules 4.690 ARNm (93,9 %). Sur les 306 ARNm restants (6,1 %), les règles « avance » et « retard » sont associées à 209 ARNm (68,3 % des 306 ARNm). Pour l'ensemble des 5 autres règles, le nombre d'ARNm est réparti de façon semblable. La répartition des ARNm entre ceux avec une avance et ceux avec un retard est différente, avec quasiment le double d'ARNm qui suivent la règle « avance » comparé à ceux qui suivent la règle « retard ». Notons que le couple « avance/retard » est le seul couple qui présente une distribution asymétrique, en faveur d'un signal plus précoce pour les embryons asexués.

Identification et classification des microARN matures

De la même façon que pour les ARNm, la discrétisation a été faite sur les microARN matures (description partie 3.1.2) et le résultat est présentée Tableau 3.5.

	sexué			asexué		
valeur	T_0/T_{1S}	T_{1S}/T_{2S}	T_{2S}/T_{3S}	T_0/T_{1A}	T_{1A}/T_{2A}	T_{2A}/T_{3A}
0	556	567	569	564	573	565
-1	1	0	2	0	0	2
1	16	6	2	9	0	6
total (-1/1)	17	6	4	9	0	8

Tableau 3.5 – Nombre de séquences de microARN matures associées aux valeurs discrètes -1, 0 et 1 pour l'ensemble des paires de pas de temps consécutifs des cinétiques sexuées et asexuées. Pour chacune des cinétiques, sexuée à gauche et asexuée à droite, le nombre de séquences de microARN mature possédant des expressions décroissantes (-1), constantes (0) et croissantes (1) entre chaque paire est noté. L'analyse a été faite sur l'ensemble des 573 séquences uniques de microARN matures. La dernière ligne présente la somme des lignes -1 et 1.

Le pourcentage de microARN matures dont l'expression varie entre deux stades est très faible. Les effectifs sont trop petits pour en tirer des conclusions robustes. On note que le nombre de microARN matures avec une valeur différente de 0 est supérieur dans le cas des embryons sexués et que la transition T_0/T_{1S}, comme pour les ARNm, regroupe la majorité des microARN matures différentiellement exprimés.

Comme pour les ARNm, les microARN matures présentant des cinétiques différentielles ont étés identifiés. 15 microARN matures possèdent des cinétiques différentes parmi les 573 microARN matures (2,6 %), ce qui représente un faible pourcentage par rapport à celui observé chez les ARNm (13,5 %).

Classification des microARN matures différentiellement régulés Comme pour les ARNm, les microARN matures ayant des cinétiques différentielles ont ensuite été classés afin de comparer l'ensemble des profils des développements embryonnaires sexués et asexués (partie 3.1.3). Les règles associées aux 15 microARN matures sont visibles Tableau 3.6 ainsi que leurs profils et des annotations/fonctions associées si elle sont connues.

Sur l'ensemble des 9 règles qui ont été définies, seulement 4 sont associées aux microARN matures avec une grande majorité associée à la règle « diminution » (67 %).

nom	sexué	asexué	règle	connu	annotations	citations
api-mir-281-5p	1,0,0	0,0,0	diminution	oui	régule l'ecdysone	[133]
api-mir-263a-5p	1,0,0	0,0,0	diminution	oui	apoptose, exprimé dans la tête	[134, 56]
api-mir-3019-5p	1,0,0	0,0,0	diminution	non		
api-mir-278-5p	0,1,1	0,0,1	diminution	oui	homéostasie énergétique, division cellules souches lignées germinales	[135, 136]
api-mir-novel183-5p	1,1,0	1,0,0	diminution	non		
api-mir-1000-5p	1,0,0	0,0,0	diminution	oui	exprimé dans la tête	[137]
api-mir-3038-3p	1,1,0	1,0,0	diminution	non		
api-mir-87-3p	1,0,0	0,0,0	diminution	oui		
api-mir-14-3p	1,1,0	1,0,0	diminution	oui	suppression mort cellulaire, métabolisme acides gras, signalisation par l'ecdysone, production d'insuline, apoptose	[138, 139, 140, 141]
api-mir-3026-5p	1,0,0	0,0,0	diminution	non		
api-mir-34-5p	0,0,0	0,0,1	augmentation	oui	neurodégénérescence	[142]
api-mir-316-5p	0,0,0	0,0,1	augmentation	oui		
api-mir-novel146-5p	-1,0,0	0,0,0	augmentation	non		
api-mir-novel185-3p	1,1,0	1,0,1	retard	non		
api-mir-1-3p	1,1,0	0,0,1	défaut	oui	différentiation cellules musculaires	[143]

Tableau 3.6 – Description des microARN matures ayant des cinétiques différentielles avec leurs profils et la règle associée ainsi que les annotations associées et les publications de référence. La colonne « connue » indique si le microARN mature a été identifié dans d'autres espèces.

Néanmoins l'effectif est trop faible pour tirer des conclusions sur la répartition des règles.

Neuf des 15 microARN matures ayant des cinétiques différentes ont déjà été identifiés dans d'autres espèces et on peut noter que api-mir-14 a déjà été identifié par Legeai *et al.* [81] comme différentiellement exprimé entre les morphes sexupares et virginopares. Si l'on observe les différentes annotations disponibles associées à ces microARN, certaines sont communes à plusieurs microARN : on peut voir que api-mir-281 et api-mir-14 sont décrits comme étant impliqués dans des processus qui concernent l'ecdysone, une des hormones chez les insectes régulant la mue et le développement des insectes. Notons que api-mir-263a et api-mir-1000 ont été détectés tous les deux comme exprimés respectivement dans les têtes du ver à soie et de l'abeille, et que api-mir-263a et api-mir-14 sont impliqués dans l'apoptose. Le microARN api-mir-278 est impliqué dans la division des cellules souches germinales, ce qui est cohérent avec le système biologique étudié ici puisque les différents embryons développent des lignées germinales soit a-méiotiques (reproduction asexuée) soit méiotique (reproduction sexuée). Le microARN api-mir-14 régule l'insuline, molécule potentiellement impliquée dans la réception de la photopériode [74]. Enfin, api-mir-1 est impliqué dans la différenciation des cellules musculaires et les cinétiques sur lesquelles nous travaillons concernent des embryons en développement et probablement la construction des muscles du futur organisme. Ainsi, plusieurs des annotations des microARN régulés lors du changement du mode de reproduction correspondent à des attendus quant aux fonctions biologiques associées.

Il nous semble important de préciser la transition entre le T_0 et les deux T_{1A} et T_{1S}, car c'est à ce stade développemental que l'on passe d'embryons flexibles (ou plastiques à la condition environnementale) à des embryons engagés dans des voies développementales. Sept microARN matures ont une différence entre T_0 et $T_{1S/1A}$, associés à la règle « défaut ». Ainsi, près de la moitié des microARN régulés le sont déjà dès cette transition précoce.

En conclusion de cette partie, nous avons identifié des catalogues d'ARNm et de microARN régulés entre les deux embryogenèses sexuées et asexuées, et classés les ARNm et microARN en types de profils d'expression (règles) les différenciant entre ces deux types d'embryogenèse.

3.2.2 Réseau d'interaction entre microARN et ARNm régulés

Une fois les ARNm et microARN matures différentiellement régulés entre les deux embryogenèses identifiés, le réseau d'interactions entre microARN matures et ARNm défini précédemment sur l'ensemble des ARNm et des microARN du puceron du pois peut être réduit aux seules interactions impliquant ces éléments différentiellement régulés.

Le Tableau 3.7 présente le nombre d'interactions impliquant des microARN matures et des ARNm différentiellement régulés sur les quatre réseaux décrits dans la partie 2.2.3 : sans seuil, avec un seuil sur le score des sites de fixation ($2^{\text{ème}}$ ligne), avec un seuil sur le nombre de sites de fixation par couples ($3^{\text{ème}}$ ligne) et avec les deux seuils (dernière ligne).

Pour le premier réseau sans seuil et avec un site, les 15 microARN régulés sont conservés. Par contre, 33,6 % des ARNm avec des cinétiques différentes ne sont ciblés par aucun des 15 microARN matures et ils ne sont donc pas dans le(s) réseau(x). Pour les deux contraintes utilisées de façon séparée ($2^{\text{ème}}$ et $3^{\text{ème}}$ lignes du Tableau), les pourcentages de conservation en comparaison au réseau sans seuil sont semblables à ceux

seuil	sites	microARN	ARNm	sites de fixation	couples
aucun	1	15 (100 %)	3.316 (100 %)	10.576 (100 %)	6.824 (100 %)
-0.3	1	15 (100 %)	1.810 (54,6 %)	2.795 (26,4 %)	2.250 (33 %)
aucun	2	14 (93,3 %)	1.491 (45 %)	5.657 (53,5 %)	1.965 (28,8 %)
-0.3	2	8 (53,3 %)	390 (11,8 %)	940 (8,9 %)	395 (5,8 %)

Tableau 3.7 – Résultats de la réduction des prédictions de TargetScan avec ou sans seuil (« seuil »), avec 1 ou 2 sites de fixation minimum (« sites ») aux microARN matures et ARNm différentiellement régulés. Sont indiqués le nombre de microARN matures (« microARN ») et d'ARNm concernés (« ARNm »), le nombre de sites de fixation, et le nombre de couples microARN/ARNm (« couple »). Entre parenthèses : pourcentage d'éléments présents en fonction du nombre total d'éléments présents dans le réseau sans seuil.

du réseau pangénomique (voir Tableau 2.10 partie 2.2.3). Sur le réseau avec l'utilisation des deux filtres (dernière ligne du Tableau) les pourcentages du nombre de sites de fixation et de couples sont similaires à ceux précédemment observés (Tableau 2.10). À l'inverse, les pourcentages sur les microARN matures et les ARNm sont beaucoup plus faibles sur le réseau avec les deux contraintes et avec les éléments différentiellement régulés : 53,3 % (réseau des éléments régulés) contre 87,4 % (réseau de l'ensemble des éléments) pour les microARN matures et de la même façon 11,8 % contre 58 % pour les ARNm. Le fait que le pourcentage de réduction en appliquant les deux filtres soit plus fort pour les microARN matures et les ARNm que pour les sites de fixation et les couples signifie qu'il y a plus d'interactions par microARN mature et par ARNm en comparaison avec le réseau réduit avec les mêmes seuils décrit partie 2.2.3 : la combinatoire entre ces deux ensembles est plus forte.

Il a été décidé de garder le réseau obtenu en ne gardant que les sites de fixation avec un score global de TargetScan inférieur ou égale à -0.3 (2$^{\text{ème}}$ ligne du Tableau 3.7) afin de ne garder que les relations les plus fortes entre microARN matures et ARNm et aussi d'éliminer des faux positifs. Ne garder que les couples de microARN matures et d'ARNm ayant au minimum deux sites de fixation aurait par contre trop fortement limité le nombre de relations et aurait donc diminué le potentiel de l'analyse de ce réseau. Dans le reste de la thèse, lorsqu'il est fait référence au *réseau d'interaction entre microARN matures et ARNm* chez *Acyrthosiphon pisum* sans autre précision, cela fera référence à ce réseau comprenant 15 microARN matures, 1.810 ARNm, 2.795 sites de fixation et 2.250 couples microARN/ARNm.

Dans la suite de ce chapitre, nous présentons de façon détaillée les annotations des ARNm présents dans le réseau ainsi que les annotations par les règles des couples microARN/ARNm. Cependant, les annotations des microARN matures présents dans le réseau n'ont pas été étudiées car les 15 microARN matures ayant des cinétiques différentielles sont présents dans le réseau.

Étude des annotations des ARNm présents dans le réseau

Étude des annotations fonctionnelles Afin de commencer à extraire de la connaissance de ce premier réseau d'interactions réduit aux éléments différentiellement régulés entre les deux embryogenèses sexuées et asexuées, nous avons utilisé et analysé les an-

notations GO. Sur l'ensemble des 1.810 ARNm présents dans le réseau, 690 ARNm possèdent une annotation GO. Un enrichissement fonctionnel a été calculé à l'aide du logiciel Blast2GO (voir partie 3.1.4) pour les ARNm avec une annotation GO présents dans le réseau (690) par rapport la totalité de ceux ayant des cinétiques différentes (1.640). Aucun enrichissement avec une p-value ajustée inférieure à 5 % n'a été détecté. Cette observation est sans doute une conséquence du faible effectif des ARNm utilisés pour cette étude.

Étude des annotations par les règles de transitions Le Tableau 3.8 présente pour chacune des règles le nombre d'ARNm avec des cinétiques différentes qui est présent dans le réseau décrit 3.2.2.

règle	nombre	ratio
augmentation	973	40 %
diminution	723	32 %
avance	53	39,6 %
retard	30	40 %
apparition pic négatif	9	36 %
disparition pic négatif	9	40,9 %
disparition pic positif	6	30 %
défaut	5	26,3 %
apparition pic positif	2	18,2 %
total	1.810	36,2 %

Tableau 3.8 – Nombre d'ARNm différentiellement régulés présent dans le réseau associé à chacune des règles. Les règles ont été organisées de façon décroissante en fonction du nombre d'ARNm. Le pourcentage affiché est le pourcentage du nombre d'ARNm associé à la règle comparé au nombre initial présenté Tableau 3.4.

La distribution des ARNm associés à chacune des règles est semblable à celle observée pour l'ensemble des ARNm avec des cinétiques différentes : les règles « augmentation » et « diminution » sont les règles majoritaires (respectivement 53,8 % et 39,9 % du nombre total d'ARNm présent dans le réseau). Sur l'ensemble des règles, les ratios ne sont pas homogènes. Les règles « diminution », « disparition pic positif », « défaut » et « apparition pic positif » possèdent des ratios inférieurs au ratio du nombre total d'ARNm avec des pourcentages respectifs de 32 %, 30 %, 26,3 % et 18,2 % contre 36,2 %. Pour les règles « disparition pic positif », « défaut » et « apparition pic positif » les ratios peuvent s'expliquer par les faibles effectifs de départ, respectivement 20, 19 et 11 (voir Tableau 3.4). Les ratios des règles « augmentation » (40 %), « avance » (39,6 %) et « retard » (40 %) sont plus élevés que le ratio total, ce qui signifie que les nombres d'ARNm suivant ces règles et ciblés par au moins un microARN mature avec des cinétiques différentielles sont plus élevés que la moyenne.

Étude des annotations par les règles des couples microARN/ARNm présents dans le réseau

Afin de mieux observer les relations entre les règles induites par les couples microARN/ARNm du réseau, le Tableau 3.9 présente les couples de règles obtenus en étudiant la corrélation entre les règles suivies par les microARN matures et les ARNm en interaction dans le réseau et le nombre de couples microARN/ARNm associés aux couples de règles.

règle microARN	règle ARNm	nombre de couples	pourcentage
diminution	augmentation	990	44 %
diminution	diminution	715	31,8 %
augmentation	augmentation	185	8,2 %
augmentation	diminution	134	6 %
diminution	avance	62	2,8 %
défaut	augmentation	50	2,2 %
défaut	diminution	30	1,3 %
diminution	retard	27	1,2 %
diminution	apparition pic négatif	8	0,4 %
diminution	défaut	7	0,3 %
diminution	disparition pic négatif	7	0,3 %
augmentation	avance	6	0,3 %
diminution	disparition pic positif	5	0,2 %
augmentation	retard	5	0,2 %
retard	augmentation	3	0,1 %
défaut	avance	3	0,1 %
augmentation	disparition pic négatif	3	0,1 %
défaut	retard	2	0,1 %
diminution	apparition pic positif	2	0,1 %
retard	diminution	1	0,04 %
augmentation	défaut	1	0,04 %
augmentation	disparition pic positif	1	0,04 %
augmentation	apparition pic négatif	1	0,04 %
défaut	apparition pic négatif	1	0,04 %
défaut	défaut	1	0,04 %
total		2.250	100 %

Tableau 3.9 – Nombre de couples de microARN matures et d'ARNm suivant les couples de règles. De gauche à droite : la règle suivie par le microARN ; la règle suivie par l'ARNm ; le nombre de couples de microARN/ARNm ; le pourcentage que représente le nombre de couples par rapport au nombre de couples total.

Un grand nombre de couples de règles possèdent des effectifs très faibles : 17 règles ont un pourcentage de représentation inférieur à 1 %. La majorité des couples microARN/ARNm est associée aux couples de règles « diminution/augmentation » (44 %) et « diminution/diminution » (31,8 %) qui représentent à eux deux 1.705 couples (75,8 %). Le nombre de couples suivant les règles « diminution/augmentation » est plus important

que celui suivant les règles « diminution/diminution ». Ceci est en accord avec la répression des ARNm par les microARN matures car les microARN matures qui possèdent une cinétique asexuée en diminution en comparaison à la cinétique sexuée semblent cibler en priorité des ARNm avec une cinétique asexuée qui augmente en comparaison à la cinétique sexuée. Il n'est pas étonnant d'observer cette répartition en pourcentage de représentation car 10 microARN matures parmi les 15 (67 %) sont associés à la règle « diminution » et la quasi-totalité des ARNm présents dans le réseau sont associés aux règles « augmentation » (53,8 %) ou « diminution » (39,9 %).

Observation des degrés des éléments du réseau

Le degré d'un nœud dans un réseau est le nombre d'arêtes relié à ce nœud. La répartition des degrés des ARNm et microARN matures donne une idée de la combinatoire entre eux. La Figure 3.5 présente l'histogramme des degrés des ARNm et le degré pour chacun des microARN matures. On peut voir que la répartition des degrés chez les ARNm est très inégale avec quasi exclusivement des ARNm ciblés par un seul microARN mature (79 %). Cela signifie que la majorité des cibles des microARN matures seront des cibles spécifiques à ces microARN matures.

Le premier fait marquant en observant les degrés des microARN matures est le nombre très élevé de cibles pour api-mir-3019-5p qui cumule à lui seul 1.300 interactions (58 %). Sur l'ensemble des ARNm ciblés par api-mir-3019-5p (1.300), 976 (54 % des ARNm du réseau) ne sont ciblés que par api-mir-3019-5p. Les écarts entre les autres degrés des microARN matures se répartissent de manière plus ou moins uniforme dans la tranche de 4 à 197 ARNm. On n'observe pas de différence de degré entre les microARN matures identifiés dans d'autres espèces et ceux identifiés uniquement chez *A. pisum*.

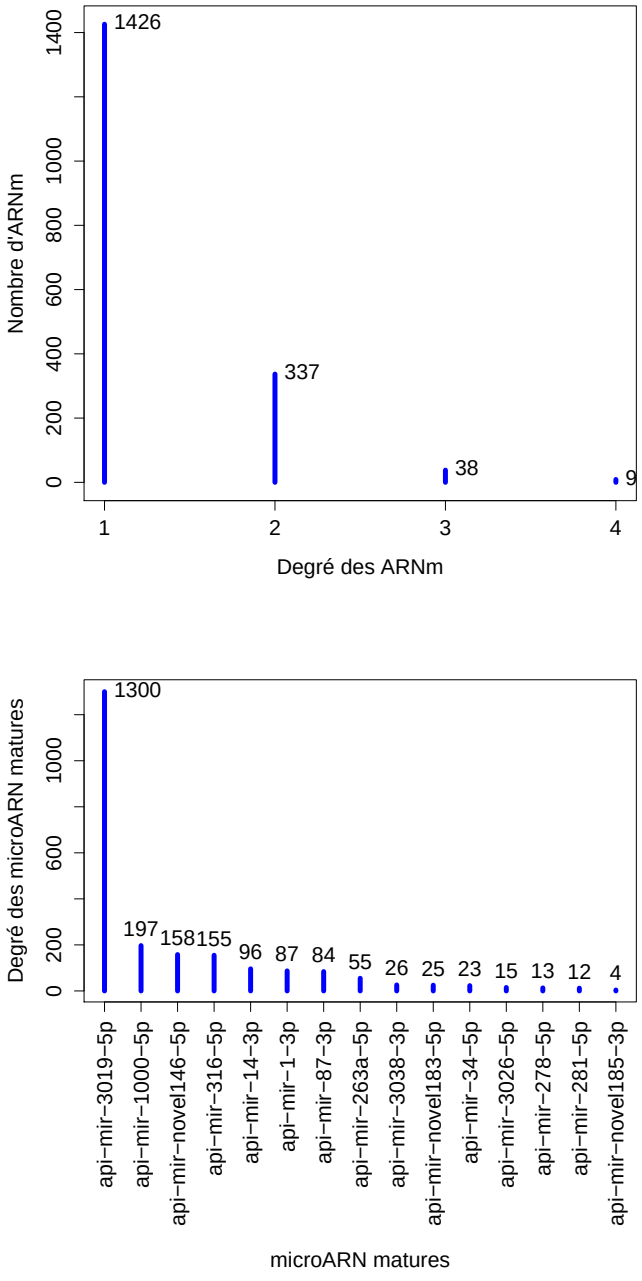


FIG. 3.5 – Histogramme des degrés des ARNm (en haut) et le degré pour chacun des 15 microARN (en bas) ayant des cinétiques différentes et présents dans le réseau.

3.3 Résumé et conclusion : des expressions d'ARNm et de microARN qui diffèrent selon le type d'embryogenèse

Dans ce chapitre, des ARNm et des microARN matures avec des cinétiques d'expression qui sont différentes selon le développement sexué ou asexué ont été identifiés. Cette identification a été conduite en discrétisant les expressions des cinétiques afin de les caractériser par une expression interprétable en termes simples vis-à-vis des mécanismes de régulations. Cette caractérisation de la variation de l'expression permet de décrire synthétiquement l'évolution des différentes expressions des ARNm et des microARN matures lors de ces deux embryogenèses. Deux sous-ensembles, l'un pour les ARNm et l'autre pour les microARN matures, ont été extraits en ne gardant que les éléments qui avaient des cinétiques discrètes différentes entre les deux morphes. Des règles ont été définies pour classer l'ensemble des différences que l'on pouvait observer entre les cinétiques. Ces règles permettent d'interpréter de façon générale les variations observées entre les deux cinétiques pour l'ensemble des éléments.

4.996 ARNm possèdent des cinétiques différentes, ce qui représente 13,5 % des ARNm d'*Acyrthosiphon pisum*. Après un enrichissement fonctionnel de ce sous-ensemble par Blast2GO et l'analyse de ces résultats par REViGO, différentes fonctions sont ressorties. Elles concernent notamment le développement, ce qui est cohérent avec l'embryogenèse, et la régulation de la transcription, ce qui impliquerait la mise en place de systèmes de régulations complexes par l'activation ou l'inactivation de gènes impliqués dans la régulation positive ou négative de la transcription. Même si lors de cette analyse il semble que la régulation négative ressorte plus que la régulation positive, il faut considérer ces résultats avec précaution car le nombre d'ARNm annotés par un terme GO ne représente que 33 % des ARNm ayant des cinétiques différentes. Nous avons pu noter que la répartition entre les règles « avance » et « retard » était différente, où deux fois plus d'ARNm suivent la règle « avance ».

Sur l'ensemble des 802 microARN matures identifiés chez *A. pisum*, seulement 15 possèdent des cinétiques différentes. Sur ces 15 microARN matures, neuf ont déjà été identifiés chez d'autres espèces, notamment *Drosophila melanogaster* ou encore *Bombyx Mori*. Sur ces microARN, 7 possèdent des annotations identifiées qui les impliquent notamment dans l'apoptose, dans la régulation par l'ecdysone, par l'insuline, la différentiation de cellules musculaires, division des cellules souches des lignées germinales ou encore leur expression dans la tête. De plus, huit microARN matures peuvent être associés à une différence d'expression entre le stade 17 (T_0) et 18 (T_1) du développement, stade où l'embryon passe d'un état flexible à un état déterminé sur son futur mode de reproduction.

L'extraction des ARNm et des microARN matures à deux sous-ensembles régulés permet de réduire le réseau à ces seuls éléments. Il a été décidé de garder le réseau réduit aux éléments régulés et constitué des interactions avec un score global de TargetScan inférieur à -0.3. L'ensemble de ces réductions, sur les éléments et sur le score, donne un réseau constitué de 15 microARN matures, 1.810 ARNm, 2.795 sites de fixation et 2.250 interactions. Les ARNm du réseau possédant une annotation GO (690) ne possèdent aucun enrichissement fonctionnel. Si l'on observe la réduction du nombre d'ARNm associés aux règles entre les ARNm du réseau et l'ensemble des ARNm différentiellement régulés, les réductions sont semblables sauf pour certaines règles où la valeur absolue est

trop faible pour pouvoir conclure. Les résultats obtenus sur la répartition des couples de règles en fonction des couples microARN/ARNm définis sur les interactions sont attendus car ils suivent les répartitions des règles des ARNm et des microARN matures. Au contraire, les répartitions des degrés dans le réseau pour les ARNm et les microARN matures sont eux surprenants : très peu d'ARNm sont ciblés par plusieurs microARN matures, 21 % et 54 % ne sont ciblés que par api-mir-3019-5p, microARN mature captant à lui seul 58 % des 2.250 interactions. Mise à part ce microARN, les degrés des autres microARN, sont uniformément répartis entre 197 et 4, possèdent des écarts plus faibles.

L'application d'une méthode d'identification de cinétiques différentes par discrétisation et la classification de ces cinétiques a permis de définir un réseau de régulation microARN/ARNm spécifique au caractère biologique étudié, le polyphénisme de reproduction chez le puceron du pois. Néanmoins, le nombre toujours élevé d'interactions rend difficile une exploration manuelle et nécessite la mise au point de méthodes d'identification de modules de régulations. Dans le chapitre 4, l'analyse de concept formel est utilisé pour réparer et visualiser le réseau et dans le chapitre 5, l'analyse de concept formel est utilisée afin d'analyser la combinatoire des interactions induites par le réseau et d'aider à la proposition de nouvelles hypothèses de travail.

Chapitre 4

Application de l'analyse de concepts formels à un réseau d'interactions microARN/ARNm

Les chapitres précédents ont présenté la création d'un réseau microARN/ARNm appliqué à une question biologique. Ce chapitre illustre comment nous avons utilisé l'analyse de concepts formels et ses extensions pour améliorer et permettre une visualisation de ce réseau. Dans un premier temps on se repose sur la structuration en concepts du réseau d'interactions pour essayer de réparer en partie les erreurs dans les prédictions. Dans un deuxième temps, on utilise la classification induite par l'analyse de concepts formels pour générer une visualisation synthétique du réseau.

4.1 Méthode de réparation de contexte formel bruité

4.1.1 L'effet du bruit sur l'analyse de concepts formels (ACF)

L'ACF est une puissante méthode d'analyse pour des données binaires. Elle permet d'extraire l'ensemble des groupes d'objets et d'attributs en relation complète. Néanmoins, cet avantage peut devenir un inconvénient dans le cas de données bruitées dû à sa sensibilité à l'absence ou à la présence ne serait-ce que d'une relation entre un objet et un attribut lors de l'inférence des concepts.

Des études ont déjà été menées sur l'analyse de concepts formels tolérante aux erreurs ou sur l'étude de concepts formels approchés [144, 145, 85]. Elles consistent principalement à retrouver dans une matrice binaire des rectangles denses de relations, c'est-à-dire autoriser certains objets à ne pas être en relation avec certains attributs à l'intérieur des rectangles. La contrainte consistant à avoir un ensemble complet de relations peut être relâchée en cherchant à maximiser le nombre de relations. Peu de travaux cherchent à récupérer des concepts formels d'origine, c'est-à-dire établis sur des données non bruitées, à partir de concepts formels obtenus sur des données d'observations bruitées.

L'une des études les plus avancées est due à Klimushkin *et al.* [146]. Ils cherchent, dans un treillis issu d'un contexte formel bruité, à retrouver les concepts d'origines. Pour cela, ils utilisent trois indices de sélection sur les concepts : la stabilité de l'extension du concept qui représente à quel point l'extension d'un concept dépend de son intension et réciproquement pour la stabilité de son intension ; la probabilité de l'intension d'un concept qui représente la probabilité que l'intension d'un concept soit clos et réciproquement pour la stabilité de son extension ; l'indice de séparation d'un concept qui indique si le concept permet de bien séparer les objets de son extension du reste des objets et de même pour son intension. Ils bruitent des contextes de deux façons différentes, soit en modifiant la valeur des cases dans le contexte avec une certaine probabilité, soit en ajoutant des objets ou des attributs au contexte. Ils montrent qu'ils peuvent récupérer une partie des concepts d'origine en utilisant l'indice de stabilité de l'intension des concepts comme filtre.

Nous commencerons en introduisant un exemple de contexte formel bruité afin d'illustrer l'effet du bruit sur l'ensemble des concepts et le treillis associé. Une analyse plus générale est développée par la suite.

Exemple de l'effet de données bruitées en analyse de concept formel

Dans le contexte $\mathbb{K}_{\text{bruité}}$ (Figure 4.1), une fausse relation (o_5, a_2) a été rajoutée au contexte $\mathbb{K}_{\text{ex}}$ (voir partie 1.4.1 et Tableau 1.1) et un score de dissimilarité a été ajouté pour l'ensemble des paires en relation qui représente l'observation de données brutes avant la transformation en tableau binaire. En ne gardant que les relations ayant un score inférieur à un seuil de -0.2, la relation d'origine (o_3, a_2) est rejetée alors que la fausse relation (o_5, a_2) est gardée. Comparé au treillis Figure 1.9, il y a maintenant 7 concepts, 3 de plus que pour le contexte $\mathbb{K}_{\text{ex}}$ (voir Figure 4.1). La suppression de la relation (o_3, a_2) a coupé le concept C_1 en deux concepts C_1' et C_1''. Le concept C_2 existe toujours dans $\mathbb{K}_{\text{bruit}}$, renommé C_2'. Deux nouveaux concepts, C_3 et C_4 ont été créés en conséquence de l'addition de la fausse relation (o_5, a_2).

	a_1	a_2	a_3	a_4
o_1	-0.3	-0.25		
o_2	-0.5	-0.45		
o_3	-0.6	-0.1		
o_4			-0.4	-0.2
o_5		-0.3	-0.28	-0.41

Tableau 4.1 – Contexte formel bruité $\mathbb{K}_{\text{bruit}}$ avec des scores de dissimilarité (la fausse relation est en rouge et la relation au dessus du seuil en bleu).

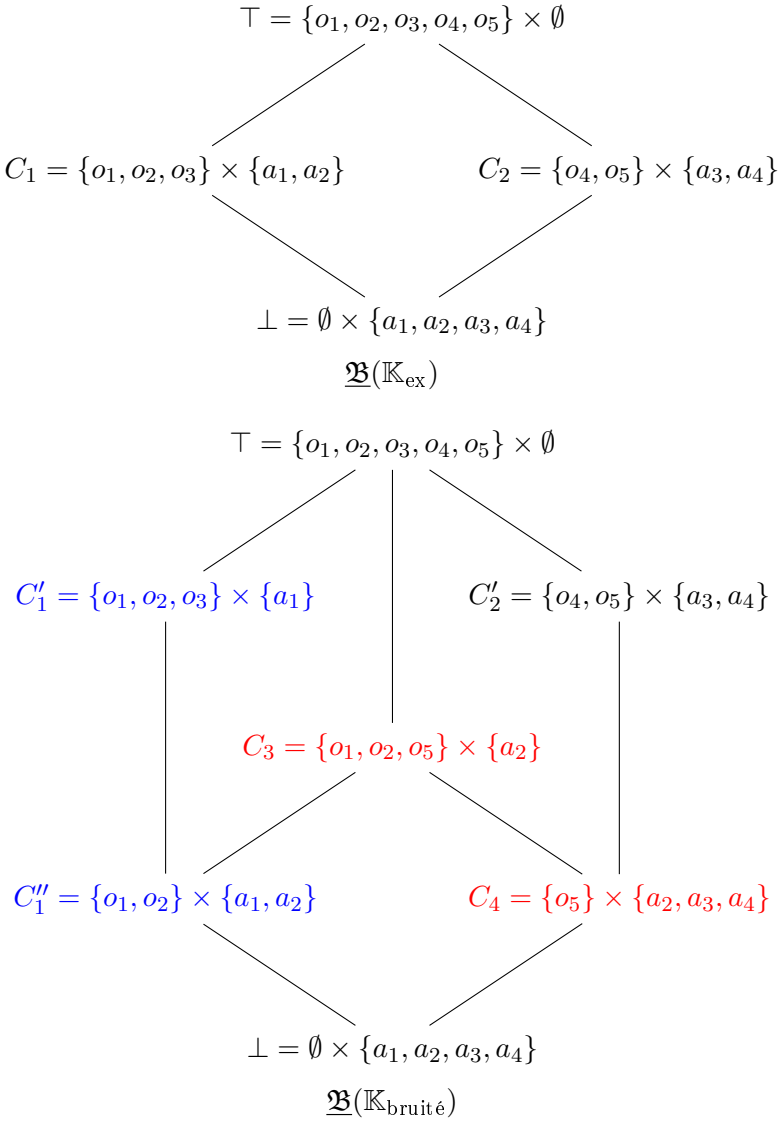


FIG. 4.1 – Le treillis de concepts $\underline{\mathfrak{B}}(\mathbb{K}_{\text{ex}})$ et le treillis $\underline{\mathfrak{B}}(\mathbb{K}_{\text{bruité}})$ associé au contexte formel bruité $\mathbb{K}_{\text{bruit}}$ en ne considérant que les relations qui possèdent un score de dissimilarité inférieur au seuil de -0.2. En bleu les concepts issus de la suppression de la relation (o_3, a_2) et en rouge les concepts issus de l'ajout de la relation (o_5, a_2).

Description du bruit : les fausses relations

Afin de mieux comprendre l'effet local de l'ajout de fausses relations sur un concept, il faut différencier deux types de relations :

- Les relations d'origine $I^o \subseteq G \times M$;
- Les fausses relations $I^f \subseteq G \times M$ avec $I^o \cap I^f = \emptyset$.

Ces deux types de relations impliquent trois types de contextes qui définissent trois types de concepts :

- Le contexte d'origine *sans fausses relations* $\mathbb{K}^o = (G, M, I^o)$ et l'ensemble des concepts d'origine $\mathfrak{C}^o$;
- Le contexte contenant *uniquement les fausses relations* $\mathbb{K}^f = (G, M, I^f)$ et l'ensemble des faux concepts $\mathfrak{C}^f$;
- Le contexte avec l'*ensemble des relations* $\mathbb{K}^{os} = (G, M, (I \cup I^s))$ et l'ensemble des concepts observé $\mathfrak{C}^{of}$.

La construction de $\mathfrak{C}^{of}$ à partir de $\mathfrak{C}^o$ et $\mathfrak{C}^f$ dépend de la contribution de chaque paire de concepts dans $\mathfrak{C}^o \times \mathfrak{C}^s$.

Considérons deux concepts $C^o = (A^o, B^o) \in \mathfrak{C}^o$ et $C^f = (A^f, B^f) \in \mathfrak{C}^f$. Comme I^o et I^f sont exclusifs, les concepts dans $\mathfrak{C}^o$ et $\mathfrak{C}^f$ sont disjoints. Cela implique que $A^o \cap A^f = \emptyset$ ou $B^o \cap B^f = \emptyset$. Supposons que $A^o \cap A^f \neq \emptyset$ et $B^o \cap B^f = \emptyset$. Alors un nouveau concept $C^{of} = (A^{of}, B^{of})$ peut être créé avec $A^{of} = A^o \cap A^f$ et $B^{of} = B^o \cup B^f$. Notons que si $A^f \subseteq A^o$ (respectivement $A^o \subseteq A^f$), alors C^f (respectivement C^o) n'est pas maximal dans $\mathbb{K}^{of}$ puisque il est inclus dans C^{of}.

Formellement, la contribution de deux concepts disjoints à l'ensemble des concepts $\mathfrak{C}^{of}$ peut être défini comme l'application d'un opérateur de fusion :

Définition 4.1 *L'*opérateur de fusion $f(.,.)$ *est défini pour une paire de concepts disjoints* $(C^i, C^j) = ((A^i, B^i), (A^j, B^j))$ *comme* $f(C^i, C^j) = \mathfrak{C}^{i \cup j}$ *où* $\mathfrak{C}^{i \cup j}$ *est l'ensemble des concepts obtenus sur les relations* $\{(A^i \times B^i) \cup (A^j \times B^j)\}$.

Les différents résultats obtenus par l'application de l'opérateur f dépendent des intersections entre les ensembles des objets et des attributs et sont énumérés ci-dessous :

$$
\begin{aligned}
f(C^i, C^j) &= \{C^i, C^j\} && \text{si } A^i \cap A^j = B^i \cap B^j = \emptyset; && (4.1)\\
&= \{(A^i \cup A^j, B^i \cup B^j)\} && \text{si } A^i = A^j \text{ ou } B^i = B^j; && (4.2)\\
&= \{C^j, (A^i \cup A^j, B^i \cup B^j)\} && \text{si } A^i \subset A^j \text{ ou } B^i \subset B^j; && (4.3)\\
&= \{C^i, C^j, (A^i \cap A^j, B^i \cup B^j)\} && \text{si } A^i \cap A^j \not\subseteq \{\emptyset, A^i, A^j\}; && (4.4)\\
&= \{C^i, C^j, (A^i \cup A^j, B^i \cap B^j)\} && \text{si } B^i \cap B^j \not\subseteq \{\emptyset, B^i, B^j\}. && (4.5)
\end{aligned}
$$

L'ensemble des concepts $\mathfrak{C}^{of}$ peut être défini comme un point fixe : $\mathfrak{C}^{of}$ est le plus petit ensemble de concepts qui couvre les concepts de $\mathfrak{C}^f$ et $\mathfrak{C}^o$ et qui est clos par f. Les concepts de $\mathfrak{C}^f$ et $\mathfrak{C}^o$ et les concepts générés par l'opérateur f appartiennent à $\mathfrak{C}^{of}$ s'ils ne sont pas couverts par d'autres concepts de $\mathfrak{C}^{of}$ comme décrit plus haut.

Description du bruit : les relations manquantes

De la même façon que pour les fausses relations, deux types de relations peuvent être distingués :

- Les relations d'origine $I^o \subseteq G \times M$;
- Les relations manquantes $I^m \subseteq I^o$.

Elles impliquent trois types de contextes qui définissent trois types de concepts :

- Le contexte d'origine *sans relations manquantes* $\mathbb{K}^o = (G, M, I^o)$ et l'ensemble des concepts originaux $\mathfrak{C}^o$;
- Le contexte contenant *seulement les relations manquantes* $\mathbb{K}^m = (G, M, I^m)$ et l'ensemble des concepts manquants $\mathfrak{C}^m$;
- Le contexte avec l'*ensemble des relations d'origine excepté les relations manquantes* $\mathbb{K}^{om} = (G, M, (I^o \setminus I^m))$ et l'ensemble des concepts $\mathfrak{C}^{om}$.

Comme pour les fausses relations, il faut décrire comment les ensembles $\mathfrak{C}^o$ et $\mathfrak{C}^m$ sont combinés dans $\mathfrak{C}^{om}$.

Soit deux concepts $C^o = (A^o, B^o) \in \mathfrak{C}^o$ et $C^m = (A^m, B^m) \in \mathfrak{C}^m$, si $A^o \cap A^m \neq \emptyset$ et $B^o \cap B^m \neq \emptyset$, alors le concept C^o ne peut être dans $\mathfrak{C}^{om}$ étant donné qu'il inclut les relations manquantes $A^m \times B^m$. À la place, deux nouveaux concepts sont créés dans $\mathfrak{C}^{om}$, $C_1^{om} = (A^o, B^o \setminus B^m)$ et $C_2^{om} = (A^o \setminus A^m, B^o)$. À noter que si $A^o \subseteq A^m$ (respectivement $B^o \subseteq B^m$), alors seul le concept C_1^{om} sera créé (respectivement C_2^{om}).

Formellement, on peut définir la contribution de deux concepts chevauchants à l'ensemble de concepts $\mathfrak{C}^{om}$ par l'application d'un opérateur d'exclusion :

Définition 4.2 *L'*opérateur d'exclusion $e(.,.)$ *est défini pour une paire de concepts chevauchants* $(C^i, C^j) = ((A^i, B^i), (A^j, B^j))$ *comme* $e(C^i, C^j) = \mathfrak{C}^{j \setminus i}$ *où* $\mathfrak{C}^{j \setminus i}$ *est l'ensemble des concepts obtenu sur les relations* $\{(A^j \times B^j) \setminus (A^i \times B^i)\}$.

Les résultats obtenus par l'application de l'opérateur e, qui dépendent des intersections des objets et des attributs, sont énumérés ci-dessous :

$$
\begin{aligned}
e(C^i, C^j) &= C^j && \text{si } A^j \cap A^i \text{ ou } B^j \cap B^i = \emptyset; && (4.6)\\
&= \{(A^j, B^j \setminus B^i), (A^j \setminus A^i, B^j)\} && \text{si } A^j \cap A^i \neq \emptyset, B^j \cap B^i \neq \emptyset; && (4.7)\\
&= \{(A^j, B^j \setminus B^i)\} && \text{si } A^j \subseteq A^i, B^j \not\subseteq B^i; && (4.8)\\
&= \{(A^j \setminus A^i, B^j)\} && \text{si } A^j \not\subseteq A^i, B^j \subseteq B^i; && (4.9)\\
&= \emptyset && \text{si } A^j \subseteq A^i, B^j \subseteq B^i. && (4.10)
\end{aligned}
$$

L'ensemble des concepts $\mathfrak{C}^{om}$ peut être défini comme un point fixe : $\mathfrak{C}^{om}$ est le plus grand ensemble de concepts qui sont inclus dans les concepts de $\mathfrak{C}^o$ et qui est clos par e. Les concepts de $\mathfrak{C}^o$ et les concepts générés par l'opérateur e appartiennent à $\mathfrak{C}^{om}$ s'ils ne contiennent aucune relation de I^m comme décrit plus haut.

Description du bruit : effet global sur le treillis de concepts

L'étude précédente met en évidence que l'augmentation du nombre de concepts dépend du type de bruit (relation fausse ou manquante) et du nombre de concepts composés uniquement de ces relations bruitées, excepté pour les équations (4.1) et (4.6) où aucun nouveau concept n'est créé.

- Pour les fausses relations, le nombre de nouveaux concepts dans $\mathfrak{C}^{of}$ dépend du nombre n_f de concepts disjoints $C^f \in \mathfrak{C}^f$ avec uniquement un ensemble qui est chevauchant avec un concept $C^o \in \mathfrak{C}^o$ et est limité par n_f ;

– Pour les relations manquantes, le nombre de nouveaux concepts $C^{om} \in \mathfrak{C}^{om}$ localement créé à partir d'un concept $C^o \in \mathfrak{C}^o$ dépend du nombre n_m de concepts $C^m \in \mathfrak{C}^m$ qui est chevauchant avec un concept C^o et est borné par 2^{n_m}.

Globalement, le nombre de nouveaux concepts augmente linéairement avec le nombre de faux concepts et exponentiellement avec le nombre de concepts manquants.

Pour réparer un contexte $\mathbb{K}^{ofm} = I^{ofm} = (G, M, ((I^o \cup I^f) \setminus I^m))$ afin de retrouver $\mathbb{K}^o$, il faut définir de nouvelles opérations qui inversent l'effet des opérateurs f et e. Ces opérations peuvent tirer parti du fait que dans la plupart des cas, les concepts qui résultent de l'application de f ou e sont reliés dans le treillis par une relation directe ou une relation jumelle.

Pour l'opérateur f, dans l'équation (4.3) les deux concepts résultants sont ordonnés dans le treillis par la relation $\prec$. Pour les équations (4.4) et (4.5), le nouveau concept est le précurseur direct ou le successeur direct de C^i et C^j dans le treillis. Pour l'opérateur e dans l'équation (4.7), les deux nouveaux concepts sont ordonnés par la relation $\prec$. Les deux ensembles A^j et B^j du concept original peuvent être retrouvés par croisement des concepts bruités.

4.1.2 Processus de réparation

Définition des opérations de réparation

Deux nouvelles opérations, *delete* et *add*, ont été définies à partir des opérateurs f et e respectivement. Ces opérations suppriment ou ajoutent des relations basées sur l'analyse du treillis de concepts. Dans la suite, nous considérons que ces opérations s'effectuent sur des paires de concepts (X, Y) avec $X = (A, B)$ et $Y = (C, D)$.

Deux types de paires de concepts peuvent être choisies :

– Les paires liées $(X, Y)_l$ ordonnées dans le treillis si $X \prec Y$ ou si $X \succ Y$;
– Les paires jumelles $(X, Y)_j$ qui possèdent un concept précédent ou suivant en commun qui diffère des éléments eux même si $\exists Z \mid X \prec Z$ et $Y \prec Z$ ou si $\exists Z \mid X \succ Z$ et $Y \succ Z$.

L'application des opérations *delete* et *add* sur les paires liées et jumelles s'effectue sur l'ensemble de la façon suivante :

$\forall\ (X, Y)_l$ ou $(X, Y)_j$:
$delete(X, Y) : \mathfrak{C} := \mathfrak{C} - Y$; $\quad Bruit := Bruit \cup (Y \setminus X)$;
$bruit(delete(X, Y)) = (Y \setminus X)$;
$\forall\ (X, Y)_l$:
Si $A \subset C$ et $D \subset B$:
$add(X, Y) : \mathfrak{C} := \mathfrak{C} - X - Y + (C, B)$; $\quad Bruit := Bruit \cup (C \setminus A) \times (B \setminus D)$;
$bruit(add(X, Y)) = (C \setminus A) \times (B \setminus D)$;
Si $C \subset A$ et $B \subset D$:
$add(X, Y) : \mathfrak{C} := \mathfrak{C} - X - Y + (A, D)$; $\quad Bruit := Bruit \cup (A \setminus C) \times (D \setminus B)$;
$bruit(add(X, Y)) = (A \setminus C) \times (D \setminus B)$.

où $\mathfrak{C}$, initialement l'ensemble des concepts observés, est l'ensemble des concepts résultant de l'application multiple des opérations *delete* et *add* et *Bruit* est l'ensemble des relations supprimées et ajoutées respectivement par les opérations *delete* et *add*.

À noter que ces opérations ne permettent pas de réparer l'ensemble des contextes bruités. En effet, dans les cas décrits par les équations (4.2) et (4.3) le concept d'origine ne devient plus accessible par ces opérations car le faux concept ajoute un ensemble d'objets ou d'attributs au concept. Dans les équations (4.8), (4.9) et (4.10), une partie de l'extension et/ou de l'intension du concept d'origine disparaît. Dans ces cinq cas, il se peut que les opérations définies ici ne permettent pas de retrouver les relations d'origine.

Dans la Figure 4.1, la sélection de la paire $(C'_2, C_4)_l$ et l'opération $delete(C'_2, C_4) = delete((\{o_4, o_5\}, \{a_3, a_4\}), (\{o_5\}, \{a_2, a_3, a_4\}))$ suppriment une fausse relation :

$$\mathfrak{C} := \mathfrak{C} - C_4 \text{ et } Bruit := Bruit \cup \{(o_5, a_2)\}.$$

De la même façon, la sélection de la paire $(C''_1, C'_1)_l$ et l'opération $add(C''_1, C'_1) = add((\{o_1, o_2\}, \{a_1, a_2\}), (\{o_1, o_2, o_3\}, \{a_1\}))$ ajoutent une relation manquante :

$$\mathfrak{C} := \mathfrak{C} - C''_1 - C'_1 + (\{o_1, o_2, o_3\}, \{a_1, a_2\}) \text{ et } Bruit := Bruit \cup \{(o_3, a_2)\}.$$

Pour une paire jumelle, la sélection de la paire $(C'_1, C_3)_j$ et l'opération $delete(C'_1, C_3) = delete((\{o_1, o_2, o_3\}, \{a_1\}), (\{o_1, o_2, o_5\}, \{a_2\}))$ suppriment une fausse relation mais aussi deux vrais relations :

$$\mathfrak{C} := \mathfrak{C} - C_3 \text{ et } Bruit := Bruit \cup \{(o_1, a_2), (o_2, a_2), (o_5, a_2)\}.$$

Définition des contraintes sur le choix des opérations

La réparation du réseau consiste à appliquer simultanément un ensemble d'opérations *delete* et *add* sur un sous-ensemble des paires de concepts issu de l'ensemble des concepts observés initialement. L'ensemble des opérations choisies est soumis à un ensemble de contraintes (où X, Y, Z et W sont des concepts et o un objet et a un attribut) :

1. $\forall X, Y, Z\ (\neg delete(X, Y) \vee \neg delete(Z, X))$: un concept utilisé pour définir la suppression d'un autre concept par une opération *delete* ne peut pas être lui aussi supprimé par une autre opération *delete* ;
2. $\forall X, Y, Z\ X \neq Y \Rightarrow (\neg delete(X, Z) \vee \neg delete(Y, Z))$: un concept ne peut pas être supprimé par deux opérations *delete* différentes ;
3. $\forall X, Y, Z\ delete(X, Y) \Rightarrow \neg add(Z, Y) \wedge \neg add(Y, Z)$: un concept supprimé par une opération *delete* ne peut pas faire aussi partie d'une opération *add* ;
4. $\forall X, Y, W, o, a\ \exists Z\ (o, a) \in bruit(delete(X, Y)) \wedge (o, a) \in bruit(delete(Z, W)) \wedge delete(X, Y) \Rightarrow delete(Z, W)$: la suppression d'une paire incluse dans un concept par une opération *delete* doit aussi être supprimée dans les concepts qui incluent cette relation ;
5. $\forall X, Y, W, o, a\ \exists Z\ (o, a) \in bruit(add(X, Y)) \wedge (o, a) \in bruit(add(Z, W)) \wedge add(X, Y) \Rightarrow add(Z, W)$: l'ajout d'une paire par une opération *add* doit aussi être ajoutée par les autres opérations *add* qui permettent d'ajouter cette relation ;

Notons que les contraintes 4 et 5 impliquent qu'une opération donnée qui nécessiterait une propagation impossible à effectuer ne sera effectivement pas appliquée.

Ces contraintes réduisent les ensembles d'opérations possibles mais il reste encore de nombreuses possibilités menant à différents résultats d'ensembles de concepts. Afin de sélectionner les ensembles les plus intéressants, nous posons le problème comme un problème d'optimisation. Pour cela nous devons définir un score associé à chaque ensemble de concepts.

Optimisation par longueur de description minimale

Les solutions obtenues par l'application simultanée des opérations *delete* et *add* sont évaluées par un score défini selon le principe de la longueur de description minimale [147]. Étant donné un ensemble de concepts $\mathfrak{C}$, on pose :

$$score(\mathfrak{C}) = \sum_{(A,B)\in\mathfrak{C}} (|A| + |B|) + \alpha\ |Bruit|\,;$$

où le paramètre α est un nombre rationnel positif (à 1 par défaut).

Ce score est minimisé pour l'ensemble des applications possibles des opérations *delete* et *add* sur les concepts de $\mathfrak{C}$ qui respectent les contraintes définies précédemment.

Nous avons modélisé l'ensemble des contraintes et la recherche d'un ensemble d'opérations produisant un ensemble de concepts de score maximum en ASP. Le programme a été testé avec la suite logiciel Potassco [93]. Ce programme prend en entrée l'ensemble des relations, des concepts et le treillis associé et renvoie en sortie l'ensemble des relations supprimées ou ajoutées. Il se déroule en cinq étapes principales :

1. Initialisation : prétraitement des données brutes pour obtenir les faits nécessaires pour la suite du programme ;
2. Sélection des concepts : définition des paires de concepts liées et jumelles ;
3. Calcul du bruit : calcul pour chacune des paires des relations supprimées ou ajoutées par les opérations *delete* et *add*. Définition des couples pour lesquels on ne peut choisir l'une des opérations, c'est-à-dire que cette opération implique une propagation qui ne peut pas être faite ;
4. Sélection des couples : sélection des opérations *delete* et *add* à effectuer et filtrage des possibilités par les contraintes 1, 2, 3, 4 et 5 ;
5. Optimisation : calcul du score et maximisation de ce score.

Finalement les relations supprimées et ajoutées sont affichées en sortie.

Afin de tester l'efficacité des opérations et du score définis pour réparer un contexte formel bruité, la méthode a été testée sur un ensemble de contextes bruités simulés.

4.1.3 Expérimentation sur des contextes bruités simulés

Plusieurs contextes aléatoires ont été simulés avec un nombre fixe d'objets et d'attributs (20, 40 ou 60) afin de tester la détection des relations fausses et manquantes. Pour chacun de ces contextes, 5 ensembles de relations ont été créés correspondant à 5 concepts avec des tailles d'extensions et d'intentions aléatoires suivant une distribution

normale de moyenne 5 et d'écart-type 2. Les contextes sont ensuite bruités avec une probabilité p_f (0,01, 0,05 ou 0,1) pour qu'une relation apparaisse entre un objet et un attribut (une fausse relation) et une probabilité p_m (0,15, 0,25 ou 0,35) pour qu'une relation disparaisse entre un objet et un attribut (une relation manquante). De plus, trois valeurs du paramètre α ont été testées (1, 1,5 ou 2). Pour chacun des jeux de paramètres (taille des contextes, niveau de bruit et paramètre α) 1.000 contextes aléatoires ont été générés. La méthode de réparation a été utilisée sur chacun de ces contextes et les résultats sont résumés dans le Tableau 4.2.

obj	att	p_f	p_m	α	*delete*				*add*			
					originale (%)		fausse (%)		non manquante (nbr)		manquante (%)	
					μ	σ	μ	σ	μ	σ	μ	σ
20	20	0,05	0,25	1	0,4	1,1	*16*	6	27,4	10,6	**37,3**	12,6
				1,5	0,1	0,3	3,2	2,6	2,9	2,3	13,8	7,9
				2	0	0	0	0	0	0	0	0
40	40	0,01	0,25	1	**2,9**	3,4	**30,7**	**12,3**	**4,9**	3,7	32,4	13,5
				1,5	1,1	1,7	**16,5**	11,4	0,8	1	13,7	7,8
				2	0	0	0	0	0	0	0	0
		0,05	0,15	1	0,3	0,9	16,1	5,9	27,9	11,3	**47,3**	16,1
				1,5	0	0,3	3,8	2,9	4	2,9	24,8	11,7
				2	0	0	0	0	0	0	0	0
			0,25	1	0,5	1,3	**16,5**	**6**	25,7	10,4	**36,5**	13,3
				1,5	0,1	0,4	3,4	2,9	2,8	2,3	14	8,7
				2	0	0	0	0	0	0	0	0
			0,35	1	0,9	1,8	16,6	5,7	23,6	9,9	**28,2**	10,5
				1,5	0,1	0,4	2,9	2,5	1,6	1,7	6,3	5,6
				2	0	0	0	0	0	0	0	0
		0,1	0,25	1	0	0,1	**1,7**	*1,9*	**83,3**	18,9	42,3	12,9
				1,5	0	0	0,2	0,5	6,3	4,3	15,9	9,2
				2	0	0	0	0	0	0	0	0
60	60	0,05	0,25	1	0,1	0,6	*9,5*	4,7	**52,1**	16,2	**35,4**	11,5
				1,5	0	0,1	0,7	0,9	2,7	2,3	12,2	7,6
				2	0	0	0	0	0	0	0	0

Tableau 4.2 – Résultat de la méthode de réparation sur des contextes bruités simulés pour des nombres différents d'objets (« obj ») et d'attributs (« att ») et différentes probabilités de fausses relations (« p_f ») et de relations manquantes (« p_m »). Pour l'opération *delete*, les moyennes (« μ ») et les écart-types (« σ ») des pourcentages (%) de fausses relations et de relations originales supprimées sur 1.000 tirages aléatoires sont indiqués. Pour l'opération *add*, la moyenne et l'écart-type du nombre (« nbr ») de relations ajoutées qui ne sont pas des relations manquantes est indiqué ainsi que la moyenne et l'écart-type du pourcentage de relations manquantes.

Sur l'ensemble des contextes, les pourcentages et le nombre de relations supprimées

ou ajoutées diminuent lorsque la valeur de α augmente pour être égale à zéro sur l'ensemble des expérimentations lorsque $\alpha = 2$. Ce résultat est cohérent avec la définition du score où α représente le poids des modifications apportées à l'ensemble des concepts en comparaison à la taille de la description des concepts réparés. Plus le poids d'une modification est fort, plus cette modification devient défavorable.

Sur l'ensemble des résultats, le pourcentage de relations originales supprimées est en moyenne très faible et est proche de zéro, excepté pour le contexte avec comme paramètres (obj ; att ; p_f ; p_m ; α)=(40 ; 40 ; 0,01 ; 0,25) où le pourcentage de relations originales supprimées est de 2,9 %, la valeur maximum sur toutes les expériences.

Une autre observation globale peut être faite sur le nombre de relations qui sont ajoutées par l'opération *add* mais qui ne sont pas manquantes. Ce nombre est élevé pour l'ensemble des tests effectués avec un paramètre α de 1, sauf pour le contexte de paramètres (40 ; 40 ; 0,01 ; 0,25) où il n'y a en moyenne que 4,9 fausses relations qui sont ajoutées. Ce nombre est très élevé pour les contextes de paramètres (40 ; 40 ; 0,1 ; 0,25) et (60 ; 60 ; 0,05 ; 0,25) (respectivement 83,3 et 52,1).

Les résultats de la réparation avec un paramètre $\alpha = 1$ des contextes de tailles 20×20 et 40×40 avec les mêmes paramètres de bruits $p_f = 0.05$ et $p_m = 0.25$ sont très semblables et ce pour les deux opérations *delete* et *add*. Environ 16 % des fausses relations sont supprimées et 37 % des relations manquantes sont retrouvées. Pour le contexte de taille 60×60 avec les mêmes paramètres, les résultats d'ajouts de relations manquantes sont quasi identiques mais le pourcentage de fausses relations supprimées (9,5 %) est plus faible.

La robustesse de la méthode avec un paramètre $\alpha = 1$ face à la probabilité d'apparition d'une fausse relation peut être étudiée en comparant les contextes de taille 40×40 et $p_m = 0,25$ avec une probabilité p_f qui varie entre 0,01, 0,05 ou 0,1. On peut voir que la moyenne du pourcentage de fausses relations supprimées diminue lorsque p_f augmente, respectivement 30,7 %, 16,5 % et 1,7 %, ainsi que l'écart-type, respectivement 12,3 %, 6 %, 1,9 %. Cette observation peut être expliquée par le fait que plus il y a de fausses relations, plus la probabilité que le concept d'origine disparaisse complètement augmente, comme décrit dans les équations (4.2) et (4.3) partie 4.1.1. Le pourcentage moyen de relations manquantes ajoutées augmente avec une valeur de p_f qui augmente. Néanmoins, le nombre de relations manquantes ajoutées est quasi identique pour les trois contextes (10, 11 et 12,4, résultats non montrés) et le nombre total de relations manquantes étant faible, une petite variation dans le nombre de relations manquantes identifiées peut entraîner une forte différence dans les pourcentages.

Pour l'étude de la probabilité de disparition des relations originelles (création de relations manquantes) avec un paramètre $\alpha = 1$, trois contextes ont été générés avec trois paramètres p_m différents : 0,15, 0,25 et 0,35 et comme paramètres fixes la taille (40×40) et la probabilité d'apparition de fausses relations ($p_f = 0,05$). Pour ces trois contextes, le nombre de relations manquantes retrouvées diminue avec la valeur de p_m qui augmente (respectivement 47,3 %, 36,5 % et 28,2 %). De la même façon que pour les résultats sur le paramètre p_f ci-dessus, augmenter la valeur de p_m revient à augmenter la probabilité de se retrouver dans les cas décrits par les équations (4.8), (4.9) et (4.10) partie 4.1.1 où le concept d'origine a complètement disparu et où il ne peut être retrouvé en croisant les concepts observés.

Une méthode de réparation de contextes bruités a été développée ici. Cette méthode aide à reconstruire des concepts qui auraient été modifiés par du bruit en s'aidant de la structure du treillis. Elle s'applique sur des matrices binaires ou pouvant être assimilées à des matrices binaires, comme par exemple avec l'utilisation d'un seuil. Dans la suite de cette partie, la méthode de réparation est utilisée sur des simulations de réseaux d'interaction microARN/ARNm.

4.1.4 Expérimentation sur des données de réseaux biologiques simulés

Il a déjà été montré qu'un réseau d'interactions entre microARN matures et ARNm peut être vu comme un contexte formel une fois les interactions microARN/ARNm discrétisées par l'utilisation d'un seuil. Dans la partie précédente, il a été montré que des modules bruités (les concepts) et leurs relations (le treillis) pouvaient aider à détecter des relations fausses ou manquantes. Ici nous faisons l'hypothèse que cette méthode est applicable à la détection des fausses prédictions induites par les méthodes de prédiction d'interactions microARN/ARNm ainsi que les vraies interactions supprimées par l'utilisation d'un seuil. L'idée est donc de récupérer une partie des erreurs de prédiction en étudiant les concepts formels et le treillis obtenus sur le réseau d'interactions. Pour tester cette idée, des réseaux sont tout d'abord simulés à partir des paramètres obtenus sur le réseau d'*Acyrthosiphon pisum* puis la méthode est appliquée sur ces réseaux.

Simulation de réseaux d'interactions microARN/ARNm issus de celui prédit chez *Acyrthosiphon pisum* Sur l'ensemble du réseau d'interaction sans seuil et réduit aux éléments avec des cinétiques différentes lors des embryogenèses, seul les 12 microARN matures possédant les degrés les plus faibles sont gardés afin de faciliter l'analyse et la simulation du réseau. De plus, pour chacun des couples microARN/ARNm, seul le site de fixation avec le plus faible score global TargetScan est gardé car c'est lui qui détermine pour quel seuil le couple sera présent ou non (en ne mettant aucune contrainte sur le nombre de sites de fixation). La distribution du score des sites de fixation pour ce réseau limité à 12 microARN matures et avec les sites de fixation aux scores les plus faible est présentée Figure 4.2. Cette distribution peut être vue comme un mélange de deux lois gaussiennes. On sait que le nombre de fausses prédictions lors de la détection des sites de fixation est élevé et que, pour la méthode TargetScan, plus le score est faible plus l'efficacité de la répression est élevée. À partir de ces constatations, on peut émettre l'hypothèse que les deux gaussiennes observées sur la distribution des scores Figure 4.2 représentent deux distributions de scores : l'une, centrée sur des faibles valeurs de scores de moyenne -0,46, représente les scores associés aux vraies prédictions et la seconde, gaussienne centrée sur des scores plus élevés de moyenne -0.2, représente les fausses prédictions.

Afin de capturer les caractéristiques de ce réseau, plusieurs paramètres ont été extraits : le nombre de microARN matures, d'ARNm et d'interactions, les deux distribution de scores en utilisant un modèle de mélanges gaussiens et les distributions des degrés pour les microARN matures à la fois pour le réseau limité aux scores très faibles et à la fois pour le réseau limité aux scores très élevés.

Ce réseau comporte 12 microARN matures, 1.468 ARNm et 2.140 interactions (le nombre de sites et d'interactions est identique car un seul site est conservé par couple).

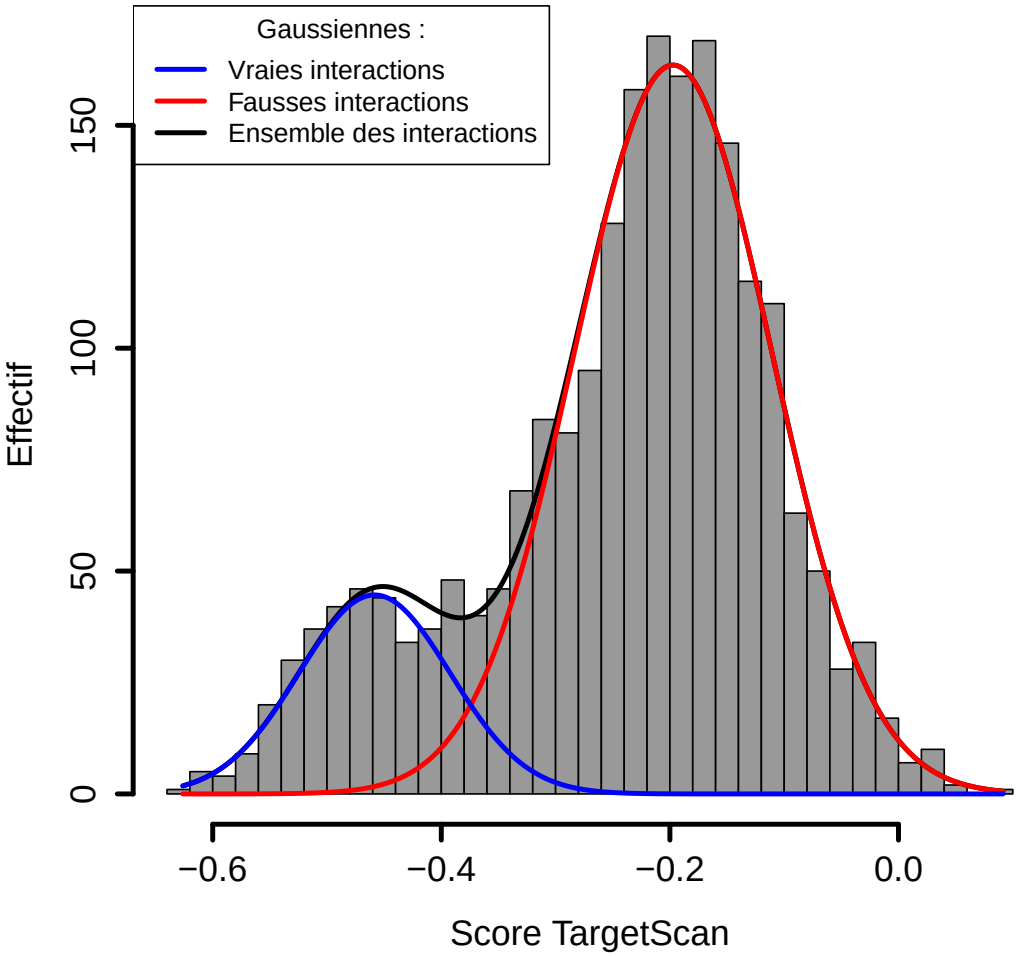


FIG. 4.2 – Histogramme des scores des sites de fixation pour le réseau limité à 12 microARN matures et avec les sites de fixation aux scores les plus faible.

Le modèle de mélange de lois gaussiennes a été obtenu à l'aide de la fonction `densityMclust` du paquet `mclust` [148, 149] du logiciel R. La méthode implémentée par cette fonction permet d'obtenir le modèle qui maximise le critère d'information bayésien [150] (BIC), critère qui dépend du maximum de vraisemblance du modèle et qui est pénalisé par le nombre de paramètres du modèle (ici le nombre lois gaussiennes). Le modèle avec la plus grande valeur BIC est un modèle à deux composantes :

1. Une moyenne de -0,46, une variance de 0,0044 et une proportion de 0,17 pour la première loi normale, que nous appellerons $\mathcal{N}_1$ et qui est considérée comme la loi normale correspondante aux vraies prédictions ;
2. Une moyenne de -0,20, une variance de 0,0074 et une proportion de 0,83 pour la seconde, que nous appellerons $\mathcal{N}_2$ et qui est considérée comme la loi normale

correspondante aux fausses prédictions.

En plus de la distribution des scores, les distributions des degrés des 12 microARN matures ont été observées pour les deux extrémités de la distribution des scores. Les deux distributions ont été extraites de la façon suivante :

1. Pour la distribution des degrés D_1 avec de faibles valeurs de scores : l'ensemble des couples de microARN/ARNm avec des scores minimum, tels que $P(X \leq s) = 0,05$ sous $\mathcal{N}_2$;
2. Pour la distribution des degrés D_2 avec de fortes valeurs de scores : l'ensemble des couples de microARN/ARNm avec des scores maximum, tels que $P(X \geq s) = 0,05$ sur $\mathcal{N}_1$.

Pour la première distribution des degrés sur les 12 microARN matures, ceux-ci varient entre 2 (1 %) et 64 (30 %) pour un nombre moyen d'interactions de 21. Pour la deuxième distribution des degrés, ils varient entre 43 (3 %) et 221 (15 %) pour un nombre moyen d'interactions de 1.511.

Les paramètres suivants ont été utilisés pour la simulation du réseau :
- Le nombre n_{micro} de microARN matures ;
- Le nombre n_{arn} d'ARNm ;
- Le nombre n_{inter} d'interactions ;
- La proportion p_2 associée à $\mathcal{N}_2$;
- Deux lois normales $\mathcal{N}_1(\mu_1, \sigma_1^2)$ et $\mathcal{N}_2(\mu_2, \sigma_2^2)$;
- Les proportions minimums min_1 et min_2 et maximums max_1 et max_2 associés aux deux distributions des degrés D_1 et D_2.

La simulation d'un réseau d'interactions microARN/ARNm suivant ces paramètres suit le protocole suivant :

1. Le tirage aléatoire en moyenne de $n_{inter} \times (1-p_2)$ vraies interactions entre n_{micro} et n_{arn}, tel que les degrés des microARN matures suivent une loi uniforme comprise entre $min_1 \times n_{inter} \times (1 - p_2)$ et $max_1 \times n_{inter} \times (1 - p_2)$;
2. Le tirage aléatoire en moyenne de $n_{inter} \times p_2$ fausses interactions entre n_{micro} et n_{arn}, tel que les degrés des microARN matures suivent une loi uniforme comprise entre $min_2 \times n_{inter} \times p_2$ et $max_2 \times n_{inter} \times p_2$;
3. L'association d'un score pour les vraies interactions suivant une loi normale $\mathcal{N}_1(\mu_1, \sigma_1^2)$ et pour les fausses interactions suivant une loi normale $\mathcal{N}_2(\mu_2, \sigma_2^2)$.

Un ensemble de 1.000 réseaux a été simulé avec ce protocole et avec les paramètres obtenus sur le réseau de *Acyrthosiphon pisum* :
- $n_{micro} = 12$;
- $n_{arn} = 1.468$;
- $n_{inter} = 2.140$;
- $p_2 = 0,83$;
- $\mathcal{N}_1(-0,46, 0,0044)$ et $\mathcal{N}_2(-0,20, 0,0074)$;
- $min_1 = 0,01$, $max_1 = 0,30$ et $min_2 = 0,03$ et $max_2 = 0,15$.

Contrainte supplémentaire sur la réparation des contextes Contrairement aux contextes bruités simulés directement dans notre première expérimentation, il s'agit ici d'obtenir une relation binaire à partir d'un tableau de score et d'un seuil sur les

scores. L'ensemble des relations manquantes potentielles s'obtient comme l'ensemble des couples de microARN/ARNm ayant un score de prédiction associé supérieur au seuil et qui n'ont pas été conservés. Les couples auxquels aucun score n'a été associé sont eux considérés comme définitivement non reliés. Une contrainte sur l'opération *add* a donc été rajoutée : seul les couples pour lesquels une interaction a été prédite peuvent être ajoutés au réseau par l'opération *add*

Résultats sur les réseaux microARN/ARNm simulés La méthode de réparation de contexte avec la nouvelle contrainte a été utilisée sur les 1.000 réseaux simulés en utilisant des seuils de -0,2, -0,25, -0,3 et -0,35 pour obtenir la relation binaire. Après avoir fixé un seuil, les vraies et fausses interactions présentes et manquantes sont définis de la façon suivante :

- Vraie interaction présente : l'interaction est vraie et le score associé est en dessous ou égal au seuil ;
- Fausse interaction présente : l'interaction est fausses et le score associé est en dessous ou égal au seuil ;
- Vraie interaction manquante : l'interaction est vraie et le score associé est au-dessus du seuil ;
- Fausse interaction manquante : l'interaction est fausse et le score associé est au-dessus du seuil.

Les ensembles des vraies et des fausses interactions manquantes définissent l'ensemble des relations pouvant être ajoutées par l'opération *add*. Les résultats sur les nombres de vraies et fausses interactions supprimées par l'opération *delete* et les vraies et fausses interactions ajoutées par l'opération *add* sont résumés Tableau 4.3.

seuil	α	*delete*				*add*			
		vraie		fausse		vraie		fausse	
		μ	σ	μ	σ	μ	σ	μ	σ
-0,2	1	9,5	21,1	14,9	29,2	0	0	2,2	1,8
	1,5	0	0	0	0	0	0	0	0
	2	0	0	0	0	0	0	0	0
-0,25	1	20,9	29,8	19,4	24,3	0	0	1,5	1,4
	1,5	0	0	0	0	0	0	0	0
	2	0	0	0	0	0	0	0	0
-0,3	1	45,7	36,1	13,9	11,2	0	0	1,3	1,4
	1,5	0	0	0	0	0	0	0	0
	2	0	0	0	0	0	0	0	0
-0,35	1	57	34,1	6,9	4,1	0	0	1,1	1,2
	1,5	0	0	0	0	0	0	0	0
	2	0	0	0	0	0	0	0	0

Tableau 4.3 – Nombre moyen (« μ ») d'interactions vraies et fausses supprimées ou ajoutées par les opérations *delete* et *add* respectivement. L'écart-type « σ » est aussi indiqué.

Le paramètre α influence plus rapidement le nombre d'interactions supprimées ou ajoutées comparé aux résultats sur les contextes bruités. Ici, plus aucune modification

des interactions ne se fait lorsque $\alpha = 1,5$. Contrairement aux résultats sur les contextes bruités, aucune vraie interaction manquante n'est ajoutée, quelque soit les paramètres utilisés. Pour l'opération *delete*, le nombre moyen de fausses interactions supprimées est supérieur au nombre moyen de vraie interactions supprimées uniquement pour le seuil de -0,2. On peut aussi noter que les écart-types sont souvent très élevés, ce qui laisse supposer que les différentes topologies des réseaux simulés varient fortement et que, comme attendu, ces différentes topologies influencent grandement le résultat de la méthode de réparation de réseau.

Les résultats obtenus sur les réseaux d'interaction microARN/ARNm simulés sont moins bons que sur les contextes bruités et le nombre de mauvaises modifications apportées est quasi systématiquement supérieur au nombre de bonnes modifications. Néanmoins, comme dit précédemment et expliqué lors de la description de la méthode, la réparation du réseau se base sur une topologie initiale de réseau principalement composée de concepts formels qui sont ensuite utilisés pour déterminer les modifications apportées à ces mêmes concepts. Lors de la simulation des contextes bruités, cette topologie a été simulée en créant par défaut un ensemble de concepts. Pour la simulation d'un réseau, seule une création des répartitions différentes des degrés entre vraies et fausses interactions a été effectuée ce qui ne garantit pas une topologie constituée principalement de concepts. Afin de pouvoir tester correctement la réparation sur des réseaux microARN/ARNm, il faudrait extraire les caractéristiques des topologies autres que les degrés pour des scores faibles et forts séparément tout en gardant la simulation des scores pour les vraies et fausses interactions.

La méthode de réparation a été adaptée à la prédiction d'interaction microARN/ARNm en ajoutant une contrainte sur l'ajout des relations. Néanmoins, cette méthode ne permet pas en l'état de réparer correctement des réseaux microARN/ARNm simulés. Ces résultats peuvent venir de la structure de concepts du réseau microARN/ARNm chez *A. pisum*, ce qui rend nécessaire la visualisation de ce réseau.

La partie suivante introduit la visualisation des concepts, qui est lié à la visualisation de graphe biparti.

4.2 Visualisation du réseau par regroupement des interactions

Le nombre d'interactions microARN/ARNm présents dans le réseau est trop important pour pouvoir en fournir une visualisation claire, interprétable par un spécialiste.

Afin d'obtenir une visualisation synthétique du réseau, l'idée est de regrouper les microARN matures et les ARNm qui ont un même comportement vis-à-vis des interactions. Il s'agit donc de « conceptualiser » le réseau, au sens de l'ACF, en mettant en évidence les regroupements sous forme de concepts et en symbolisant l'ensemble des interactions qu'ils représentent par une seule arête entre les groupes constitués. On réduit ainsi fortement le nombre d'arêtes à visualiser. D'un point de vue théorique, un réseau microARN/ARNm est un graphe biparti et un concept est une biclique de taille maximale de ce graphe. Visualiser l'ensemble des interactions à l'aide de concepts revient à couvrir le graphe par un ensemble de bicliques.

La couverture d'un graphe biparti par une partie, minimum ou non, de ses bicliques maximales est un problème étudié depuis quelques années [151, 152] et qui est utilisé en bio-informatique [153]. Une des méthodes de couverture de graphe biparti par ses bicliques, Power Graph (PG), a été développée par Royer *et al.* [154]. Elle est utilisée dans différentes études sur des réseaux biologiques [154, 155, 156] ou encore sur des réseaux sociaux [157, 158].

Dans une première sous-partie la méthode sera tout d'abord décrite, puis une reconstruction rationnelle de PG sera proposée dans le cadre de l'ACF. Dans un troisième temps, nous présenterons les résultats de la réduction de la complexité de notre réseau d'interactions chez *Acyrthosiphon pisum* par PG et nous terminerons en montrant les possibilités étendues que permet l'ACF pour produire des visualisations et des analyses alternatives.

4.2.1 Description de la méthode de compression et de visualisation Power Graph

La description de la méthode est tirée de l'article [154]. Étant donné un graphe $G = (V, E)$ où V est l'ensemble des nœuds et $E \subseteq V \times V$ est l'ensemble des arêtes, un *power graph* (PG) $G' = (V', E')$ est un ensemble de *power nodes* $V' \subseteq P(V)$ regroupant des nœuds reliés par un ensemble de *power edges* $E' \subseteq V' \times V'$. Si deux power nodes sont reliés par un power edge dans G', alors tous les nœuds présents dans le premier power node sont reliés par une arête à tous les nœuds dans le deuxième power node dans G. Trois contraintes sont nécessaires pour fournir une visualisation claire du PG G' :

1. *Hiérarchie des power nodes* : deux power nodes sont soit distincts l'un de l'autre, soit l'un inclus dans l'autre ;
2. *Couverture des arêtes* : toute arête du graphe G est représentée par un power edge dans G' ;
3. *Partitionnement des power edges* : une arête du graphe G est représentée par un seul power edge dans G' ;

L'algorithme utilisé pour détecter le PG minimal d'un graphe donné recherche tout d'abord l'ensemble des power nodes possibles par un clustering hiérarchique [159] puis

ajoute les power edges par une optimisation gloutonne. Il s'agit donc d'un algorithme d'optimisation local, qui ne garantit pas de trouver le minimum global recherché.

Recherche de l'ensemble des power nodes possibles

L'ensemble des power nodes possibles est obtenu par un clustering hiérarchique en utilisant l'indice de Jaccard [120] comme mesure de similarité entre les nœuds. Pour cela, l'indice de Jaccard utilise la notion de premier voisin, ou voisin direct d'un nœud : l'ensemble des premiers voisins d'un nœud u est constitué des nœuds en interaction directe avec nœud u. L'indice de Jacquard entre deux nœuds est : $J(N_1, N_2) = (N_1 \cap N_2)/(N_1 \cup N_2)$, où N_1 est l'ensemble des premiers voisins du premier nœud et N_2 l'ensemble des premiers voisins du second nœud. L'indice de Jacquard est donc compris entre 0 (les deux nœuds n'ont aucun voisin en commun) et 1 (les deux nœuds ont les mêmes voisins).

Recherche des power edges

La génération des power edges du power graph (PG) est effectuée par une recherche gloutonne qui permet d'obtenir la solution minimale ou une approximation de cette solution [160]. Pour chaque paire de power nodes obtenue, un power edge peut être associé. À chaque itération de la recherche, le power edge qui permet de réduire au mieux le nombre d'arêtes du graphe, c'est-à-dire celle qui couvre la plus grande surface tout en vérifiant les conditions sur les power nodes et les power edges, est ajouté. Si une power edge est créée entre deux power nodes potentiel, alors ces deux power nodes sont ajoutés dans le PG. Les interactions couvertes par cette power edge sont otées du graphe et la recherche continue. À noter qu'un power node peut ne contenir qu'un seul nœud.

Disponibilité de la méthode

La méthode d'abstraction d'un graphe en PG est disponible sous la forme d'un module, CyOog, pour le logiciel Cytoscape [161], un outil de visualisation et d'analyse de graphe. Ce module est aussi disponible sur internet dans une version pré-installée[1].

4.2.2 Parallèle entre Power Graph et l'ACF

Dans le cas particulier où G est un graphe biparti, il existe un parallèle entre la modélisation et la réduction par la méthode de PG du réseau et l'ACF sur ce même réseau. En effet, un concept formel $C = (A, B)$ peut être vu comme un ensemble de deux power nodes et d'une power edge. Les deux power nodes sont constitués pour l'un des éléments de A, pour l'autre des éléments de B, et la power edge couvre les arêtes de $A \times B$. De ce fait, le treillis des concepts, donc l'ensemble des concepts, énumère l'ensemble des triplets $(PN_1, PN_2, PE_{1,2})$ maximaux où PN_1 et PN_2 sont deux power nodes potentiels et $PE_{1,2}$ le power edge potentiel entre PN_1 et PN_2. Ils sont maximaux dans le sens où la couverture en nombre d'arêtes de $PE_{1,2}$ est maximale pour PN_1 et PN_2.

[1] www.biotec.tu-dresden.de/research/schroeder/powergraphs/download-cytoscape-plugin.html

Les trois contraintes des PG, hiérarchie des power nodes, couverture des power edges et partitionnement des arêtes peuvent être transposées en terme d'ACF. On note $choix(C_i)$ le choix d'un concept $C_i = (A_i, B_i)$ représentant deux power nodes et un power edge. On note $I_i = A_i \times B_i$ l'ensemble des arêtes couvertes par ce concept et e une arête du graphe G. Les contraintes peuvent alors être définies de la façon suivante :

1. *Hiérarchie des power nodes* en ACF : $\forall C_1, C_2\ (choix(C_1) \wedge choix(C_2)) \Rightarrow (A_1 \cap A_2) \in \{\emptyset, A_1, A_2\} \wedge (B_1 \cap B_2) \in \{\emptyset, B_1, B_2\}$. Deux concepts choisis ont leurs ensembles A_1, A_2 et B_1, B_2 soit disjoints, soit l'un inclus dans l'autre ;
2. *Couverture des arêtes* en ACF : $\forall e\ \exists choix(C_1) \wedge e \in I_1$. Pour toute les arêtes e de G, il existe un concept choisi qui couvre cette arête ;
3. *Partitionnement des power edges* en ACF : $\forall C_1 \neq C_2\ (choix(C_1) \wedge choix(C_2)) \Rightarrow I_1 \cap I_2 = \emptyset$. Deux concepts choisis couvrent des arêtes différentes de G.

Les contraintes de hiérarchie et de partitionnement peuvent être exprimées comme des contraintes sur le treillis de la façon suivante :

1. *Hiérarchie des power nodes* sur le treillis : $\forall C_1, C_2\ (choix(C_1) \wedge choix(C_2)) \Rightarrow sup(C_1, C_2) \in \{C_1, C_2, \top\} \wedge inf(C_1, C_2) \in \{C_1, C_2, \bot\}$ et $sup(C_1, C_2) = \top \wedge C_1 \neq C_2 \neq \top \Rightarrow intension(\top) = \emptyset$ et $inf(C_1, C_2) = \bot \wedge C_1 \neq C_2 \neq \bot \Rightarrow extension(\bot) = \emptyset$, avec *sup* et *inf* qui renvoient respectivement la borne supérieure et la borne inférieure de deux concepts et *intension* et *extension* qui renvoient respectivement l'intension et l'extension d'un concept. Soit deux concepts ont comme borne supérieure et comme borne inférieure l'un et l'autre des deux concepts. Soit deux concepts différents choisis ont comme borne supérieure le supremum du treillis ($\top$) et comme borne inférieure l'infimum du treillis ($\bot$) et l'intension du supremum et l'extension de l'infimum sont égaux à l'ensemble vide ;
2. *Partitionnement des power edges* sur le treillis : $\forall C_1, C_2\ choix(C_1, C_2) \Rightarrow C_1 \not< C_2 \wedge C_1 \not> C_2$. Deux concepts choisis sont incomparables.

De plus, les étapes de sélection des power edges correspondent à des sélections de concepts et sous-concepts dans le contexte formel du graphe G, un sous-concept étant défini comme une biclique résultant de la suppression d'une partie des relations couvertes par un ou deux autres autre concepts. Les relations sont supprimées si elles sont couvertes par un power edge sélectionné. On retrouve dans ce processus de suppression une partie de la définition des relations manquantes décrite sur la réparation des contextes formels bruités (partie 4.1.1). On peut donc là aussi définir un opérateur afin de caractériser l'action de suppression d'un concept sur un autre concept :

Définition 4.3 *Étant donnés deux concepts* $C^i = (A^i, B^i)$ *et* $C^j = (A^j, B^j)$, *l'*opérateur de soustraction $s(.,.)$ *est défini comme* $s(C^i, C^j) = (A^i \times B^i \setminus A^j \times B^j)$ *si* $C^i < C^j \vee C^i > C^j$ *et* $s(C^i, C^j) = (A^i \times B^i)$ *sinon.*

À partir d'un concept C^i donné, l'opérateur s renvoie un rectangle constitué de la soustraction d'un concept C^j au concept initial C^i s'ils sont sur une même chaîne. Sinon l'opérateur renvoie le concept C^i initial. À l'aide de l'opérateur s, les arêtes et les nœuds déjà couverts par un concept peuvent être enlevés des autres concepts situés sur la même chaîne. Cet opérateur est similaire à l'opérateur e défini sur la réparation des contextes formels bruités. Appliquer l'opérateur s revient à appliquer l'opérateur e sur deux concepts d'une même chaîne.

On peut définir un opérateur de double soustraction comme une extension de l'opérateur de soustraction qui permet de formaliser la soustraction de deux concepts à un seul :

Définition 4.4 *Étant donnés trois concepts* $C^i = (A^i, B^i)$, $C^j = (A^j, B^j)$ *et* $C^k = (A^k, B^k)$, *l'*opérateur de double soustraction $s_2(.,.,.)$ *est défini comme* $s_2(C^i, C^j, C^k) = (A^i \times B^i \setminus (A^j \times B^j \cup A^k \times B^k))$ *si* $C^k < C^i < C^j$ *et* $s_2(C^i, C^j, C^k) = (A^i \times B^i)$ *sinon.*

Cette opérateur s_2 permet de soustraire deux concepts C^j et C^k à un concept C^i si ils sont tous sur la même chaîne, c'est-à-dire un même chemin sur le treillis.

À l'aide de ces deux opérateurs de soustraction, on peut définir une représentation inspirée par les PG. On note $valide(\mathfrak{C})$ le fait que les power nodes et power edges obtenus à partir d'un ensemble $\mathfrak{C}$ de concepts ou sous-concepts respectent les conditions de hiérarchie et de partitionnement. La stratégie suivante a été mise en place pour obtenir la réduction G' d'un graphe biparti G à l'aide de son ensemble de concept $\mathfrak{C}$ à l'aide de l'ACF :

1. Initialisation : prétraitement des données brutes pour obtenir les faits nécessaires pour la suite du programme ;
2. Sélection des concepts : sélection d'un ensemble $\mathfrak{C}_1$ des concepts tel que : $valide(\mathfrak{C}_1)$;
3. Sélection des sous-concepts par simple soustraction : sélection d'un ensemble $\mathfrak{C}_2$ de sous-concepts tel que : $valide(\mathfrak{C}_1 \cup \mathfrak{C}_2)$ et $\forall C^s \in \mathfrak{C}_2 \Rightarrow \exists C_1 \in (\mathfrak{C} \setminus \mathfrak{C}_1) \wedge \exists C_2 \in \mathfrak{C}_1 \wedge s(C_1, C_2) = C^s$. L'ensemble des concepts $C_1 \in (\mathfrak{C} \setminus \mathfrak{C}_1)$ utilisé pour obtenir $\mathfrak{C}_2$ est noté $\mathfrak{C}'_2$;
4. Sélection des sous-concepts par double soustraction : sélection d'un ensemble $\mathfrak{C}_3$ de sous-concepts tel que : $valide(\mathfrak{C}_1 \cup \mathfrak{C}_2 \cup \mathfrak{C}_3)$ et $\forall C^{s_2} \in \mathfrak{C}_3 \Rightarrow \exists C_1 \in (\mathfrak{C} \setminus (\mathfrak{C}_1 \cup \mathfrak{C}'_2)) \wedge \exists C_2, C_3 \in (\mathfrak{C}_1 \cup \mathfrak{C}_2) \wedge s_2(C_1, C_2, C_3) = C^{s_2}$;
5. Optimisation : calcul de la surface couverte par $\mathfrak{C}_1 \cup \mathfrak{C}_2 \cup \mathfrak{C}_3$ et maximisation de cette surface.

Cette méthode permet de définir un regroupement des nœuds et des arêtes basé sur les concepts et l'application des opérateurs s et s_2.

L'avantage de cette méthode est que l'optimum de l'étape 5 peut être affiné pour optimiser différents critères de regroupement. Par exemple, dans notre cas d'un réseau microARN/ARNm, on peut vouloir favoriser le regroupement des microARN ou encore favoriser le regroupement des ARNm possédant des annotations similaires. Dans la suite de la thèse, la méthode PG et notre méthode basée sur l'ACF sont appliquées au réseau microARN/ARNm chez *Acyrthosiphon pisum*. Nous n'avons pas eu le temps de développer en profondeur la méthode par ACF et les résultats présentés ici doivent donc être considérés comme préliminaires.

4.2.3 Application au réseau d'interaction microARN/ARNm chez *Acyrthosiphon pisum*

L'ensemble des visualisations a été produit en utilisant le module CyOog [154]. La méthode PG a été utilisée sur notre réseau avec les valeurs par défaut (Figure 4.3). L'extraction a permis de réduire notre réseau de 15 microARN matures, 1.810 ARNm

et 2.250 interactions à 14 power nodes regroupant des microARN matures (powerMicro), 49 power nodes regroupant des ARNm (powerARN) et 98 power edges. Le facteur de réduction apporté est donc de 95,6 % sur le nombre d'arêtes. Sur l'ensemble des 14 powerMicro, un seul est constitué de deux microARN matures. Les 13 autres incluent un microARN mature seulement. Les deux microARN matures au sein du même powerMicro sont les microARN matures api-mir-14-3p et api-mir-34-5p. Les ARNm se regroupent dans 49 powerARN et tous les powerARN incluent plusieurs ARNm. Même après la réduction du nombre d'arêtes, le réseau reste touffu surtout à cause de la taille des powerARN.

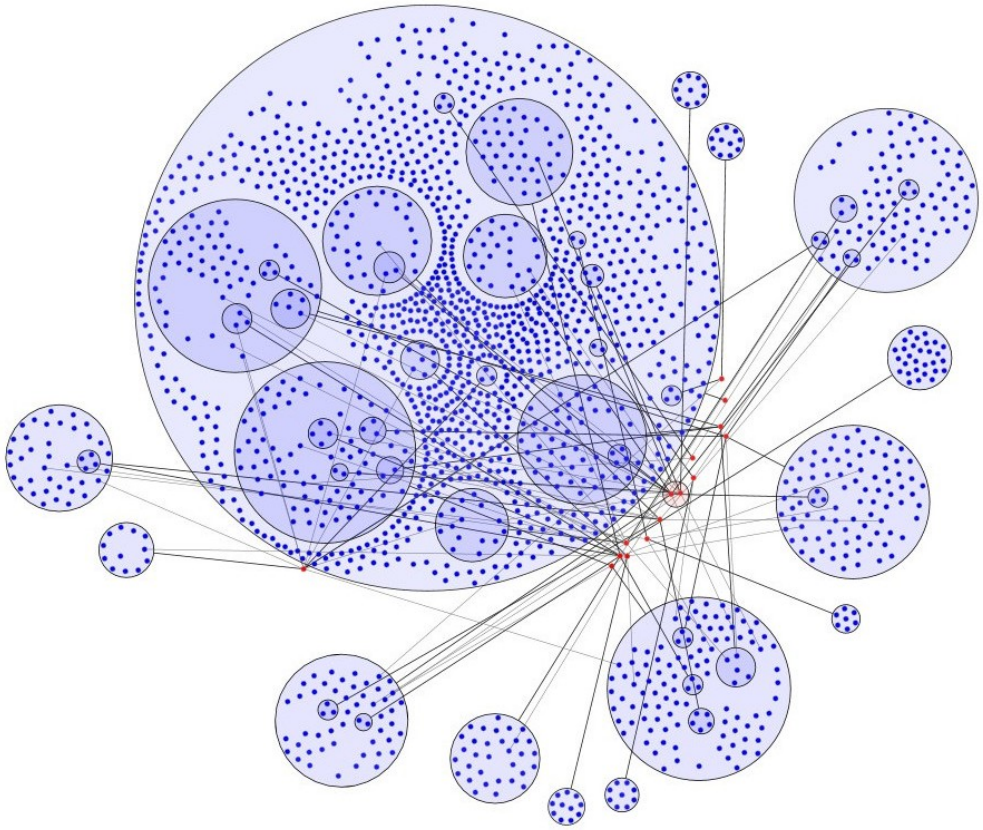


FIG. 4.3 – Power graph du graphe d'interaction entre microARN matures et ARNm avec des cinétiques différentielles chez *Acyrthosiphon pisum*. Les microARN matures et powerMicro sont en rouge et les ARNm et powerARN en bleu. Image obtenue grâce au module CyOog.

Nous avons donc décidé de supprimer les nœuds qui ne participent qu'a une seule arête, ici les microARN matures qui ciblent un seul microARN et les ARNm ciblés par un seul microARN. Ces interactions sont en effet spécifiques d'un microARN ou d'un ARNm et peuvent être simplement listées à part. Aucun microARN mature n'est supprimé et 1.426 ARNm sont supprimés, pour un nombre total d'ARNm dans le nouveau réseau de 384. Le PG obtenu sur ce nouveau réseau de 824 interactions, que nous appellerons « petit réseau » pour plus de simplicité, est présenté Figure 4.4.

Sur ce « petit réseau », on obtient 12 powerMicro, 43 powerARN et 91 power edges

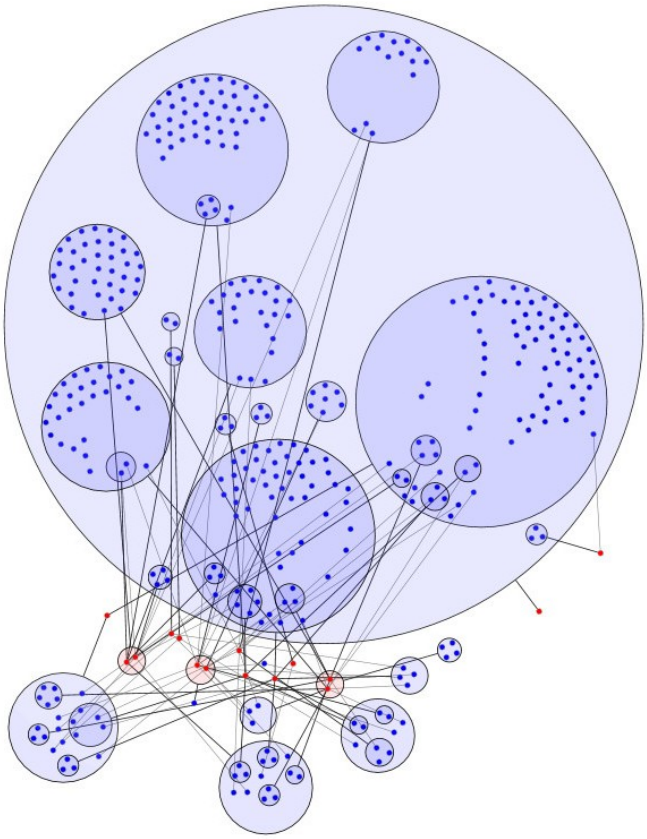


FIG. 4.4 – Power graph du « petit réseau ». Les microARN matures et powerMicro sont en rouge et les ARNm et powerARN en bleu. Image obtenue grâce au module CyOog.

pour un facteur de réduction de 90 %. Sur les 12 powerMicro, trois incluent deux microARN matures. Ces trois regroupement sont : api-mir-1-3p avec api-mir-87-3p, api-mir-3038-3p avec api-mir-novel183-5p et api-mir-14-3p avec api-mir-34-5p. Sur les 43 powerARN, 41 incluent plusieurs ARNm. Les plus grandes réductions entre le réseau d'origine et le « petit réseau » ne s'opèrent pas sur le nombre de power edges, qui passe de 98 à 91 (le facteur de réduction est en fait légèrement inférieur), mais sur la taille des powerARN ainsi qu'un meilleur regroupement des microARN matures au sein des powerMicro.

Notre méthode de visualisation basée sur l'ACF a également été appliquée au « petit réseau » (Figure 4.5). La méthode génère un ensemble de 11 powerMicro, 42 powerARN et 242 power edges pour un facteur de réduction de 70 %. Sur les 11 powerMicro, quatre incluent deux microARN matures et sur les 42 powerARN, neuf incluent plusieurs ARNm. Les microARN matures regroupés sont : api-mir-3038-3p et api-mir-novel183-5p, api-mir-novel146-5p et api-mir-3019-5p, api-mir-1-3p et api-mir-87-3p et finalement api-mir-1000-5p et api-mir-263a-5p. Une première comparaison entre cette méthode et la méthode PG est que la réduction sur les microARN matures est supérieure, quatre powerMicro contre trois qui incluent plusieurs microARN matures, mais la réduction sur les arêtes est grandement inférieure, 242 contre 91. De façon purement visuelle, il

semble que certains regroupement d'ARNm dans des powerARN pourraient encore être effectués. Ceci indique que la méthode développée avec l'ACF ne permet pour l'instant pas de récupérer l'ensemble des regroupements possibles. Ceci expliquerait le nombre plus faible de powerARN incluant plusieurs ARNm, 9 contre 41. De plus, on peut voir que les regroupements effectués sur les microARN matures sont différents : api-mir-14-3p et api-mir-34-5p ne sont plus regroupés ensemble par exemple. L'intérêt de la méthode basée sur les concepts réside dans la flexibilité de sa fonction objectif qui peut être modifiée pour optimiser d'autres critères que simplement la surface couverte.

La Figure 4.6 présente la réduction du « petit réseau » par la méthode basée sur l'ACF, mais où cette fois ci nous avons utilisé deux critères d'optimisation. Nous avons procédé de façon prioritaire à une optimisation du recouvrement des arêtes correspondantes à des couples de type « contraire » (Tableau 4.4 pour la liste des couples « contraires ») puis utilisé l'optimisation de la couverture pour départager les ex aequo. Le résultat comporte 12 powerMicro, 41 powerARN et 265 power edges. Sur les 15 powerMicro : un inclut deux microARN matures (api-mir-316-5p et api-mir-14-3p) et un deuxième inclut trois microARN matures (api-mir-novel146-5p, api-mir-3019-5p et api-mir-1000-5p). Parmi les 41 powerARN, neuf incluent plusieurs ARNm. Les regroupements obtenus sont assez différents entre les deux visualisations générées par des optimisations différentes, notamment sur les powerMicro. Une fois encore, on constate des regroupements d'ARNm qui ne sont pas supportés par des regroupements en powerARN. Néanmoins, cette figure illustre bien le fait que grâce à notre méthode et en changeant la fonction objectif, nous pouvons modifier le regroupement des éléments et générer des visualisations qui permettent de meilleurs interprétations biologiques en fonction du problème posé.

En résumé, cette partie introduit le problème d'une visualisation basée par les bicliques d'un graphe biparti. Pour répondre à ce problème, la méthode PG a été identifiée comme l'une des solutions possibles pour couvrir ce type de graphe par un partitionnement spécifique. Nous avons transposé cette approche dans le cadre de l'ACF et proposé une nouvelle méthode, inspiré par la méthode PG et basée sur l'ACF, qui permet d'étendre les possibilités de visualisation par la modification de l'optimum à atteindre.

microARN	ARNm	microARN	ARNm
diminution	augmentation	augmentation	disparition pic positif
augmentation	diminution	disparition pic positif	augmentation
avance	retard	augmentation	apparition pic négatif
retard	avance	apparition pic négatif	augmentation
diminution	avance	disparition pic négatif	apparition pic négatif
avance	diminution	disparition pic négatif	disparition pic positif
augmentation	retard	apparition pic négatif	disparition pic négatif
retard	augmentation	disparition pic positif	disparition pic négatif
diminution	disparition pic négatif	disparition pic positif	apparition pic positif
disparition pic négatif	diminution	disparition pic positif	disparition pic négatif
diminution	apparition pic positif	apparition pic positif	disparition pic positif
apparition pic positif	diminution	disparition pic négatif	disparition pic positif

Tableau 4.4 – Liste des couples règles considérées comme « contraires » en fonction de la règle suivie par le microARN mature et par l'ARNm

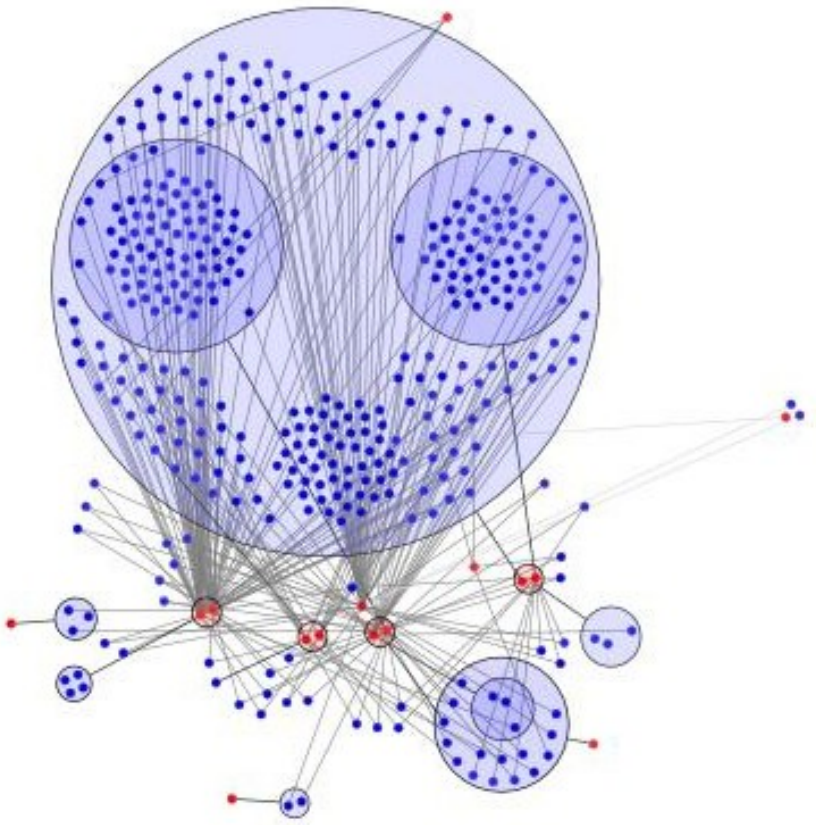


FIG. 4.5 – Réduction du « petit réseau » par la méthode basée sur l'ACF. Les microARN matures et powerMicro sont en rouge et les ARNm et powerARN en bleu. Image obtenue grâce au module CyOog.

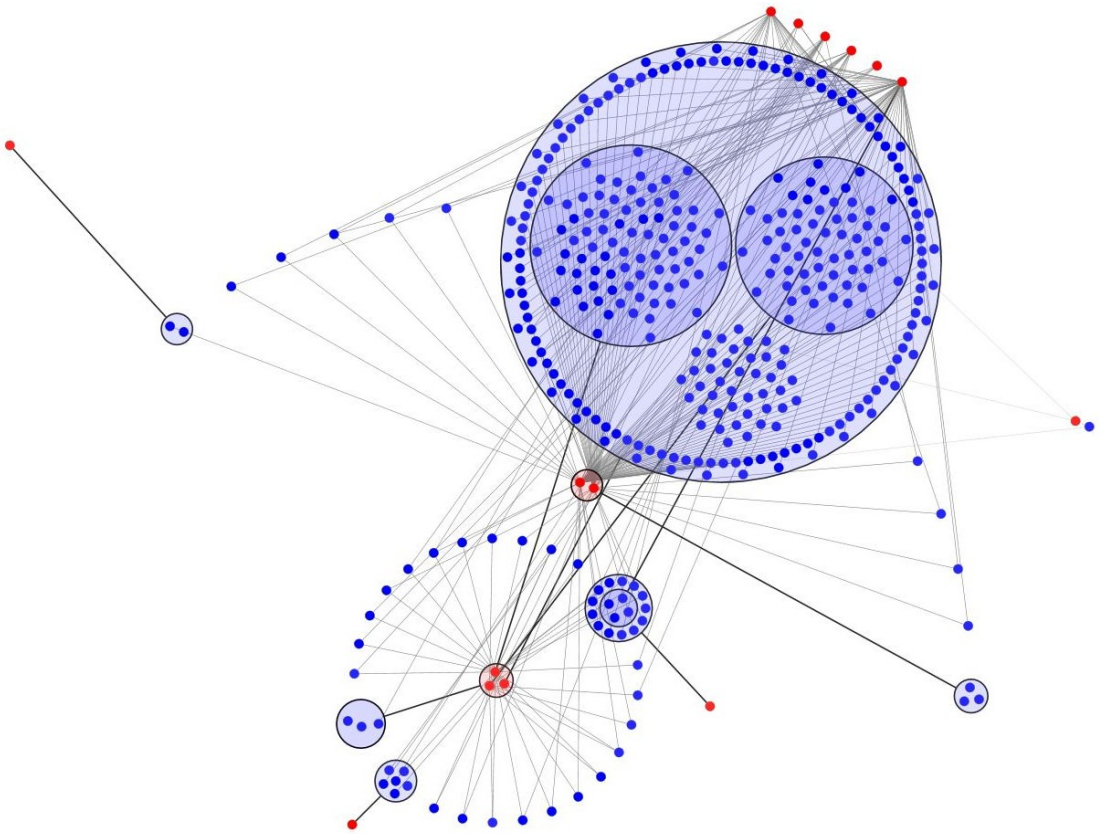


FIG. 4.6 – Réduction du « petit réseau » par la méthode basée sur l'ACF avec l'optimisation de la couverture d'arêtes « contraire ». Les microARN matures et powerMicro sont en rouge et les ARNm et powerARN en bleu. Image obtenue grâce au module CyOog.

4.3 Résumé et conclusion : l'analyse de concepts formels dans le cadre de graphes biologiques bipartis

Ce chapitre décrit deux méthodes issues de l'ACF et développées dans le cadre d'un réseau d'interactions microARN/ARNm. La première propose de réparer un graphe biparti à l'aide de l'ACF. On part de l'hypothèse qu'il existe un graphe d'origine, le vrai graphe, et que ce graphe est constitué principalement de modules d'interaction : des concepts formels. La formalisation de l'effet du bruit au sein d'un contexte formel nous a permis de détecter l'effet de ce bruit sur le nombre de concepts et le changement que ce bruit induit dans le treillis de concepts. À partir de cette formalisation, deux opérations ont été proposées pour détecter et réparer l'effet du bruit sur le treillis. Ces opérations ont été testées sur des jeux de données *in silico* simulés à la fois sans paramétrage biologique et à la fois à l'aide de paramètres dérivés du réseau microARN/ARNm chez *Acyrthosiphon pisum*. Les résultats obtenus sur les simulations sans *a priori* sont convaincants et la méthode permet de réparer un partie du réseau. Les résultats obtenus sur les graphes simulés à partir de paramètres issus du réseau réel sont moins bons. En effet, la méthode ne permet pas dans ce cas là de réparer convenablement le graphe. Néanmoins, les paramètres extraits du réseau microARN/ARNm ne sont pas forcément ceux qui permettent de reproduire convenablement la structure sous-jacente du réseau. Une version préliminaire de cette méthode a fait l'objet d'une publication qui a été sélectionnée parmi les actes de la conférence European Conference on Data Analysis 2013[2]. Cette publication est disponible en annexe 6.2.2.

La deuxième partie introduit le problème de la visualisation d'un graphe biparti. La méthode Power Graph (PG) est une solution qui permet de répondre à ce problème en couvrant les arêtes d'un graphe biparti par des bicliques par une approche hiérarchique. La méthode PG a tout d'abord été transposée dans le cadre de l'ACF, puis une deuxième méthode permettant de visualiser un graphe biparti par ses concepts et sous-concepts couvrant au mieux ce graphe a été proposée. Elle s'inspire de la méthode PG. Cette méthode nouvelle permet de minimiser le nombre d'arêtes en représentant un concept ou sous-concept par une seule arête. L'avantage de notre méthode est qu'elle permet d'implémenter différentes optimisations, autres qu'une optimisation de couverture de surfaces comme la méthode PG. Cet avantage a été illustré par l'optimisation du regroupement des arêtes associées à des couples de règles définis comme « contraire » sur la base des microARN matures et des l'ARNm associés à ces arêtes.

[2]Livre disponible fin 2014 : `http://www.springer.com/statistics/book/978-3-662-44982-0`

Chapitre 5

Étude du réseau d'interactions par l'analyse de concepts formels

Le chapitre 2 a présenté la création d'un premier réseau d'interactions entre les microARN et les ARNm chez *Acyrthosiphon pisum* et le chapitre a 3 décrit la réduction de ce réseau aux microARN et ARNm qui possèdent des cinétiques différentes entre l'embryogenèse sexuée et l'embryogenèse asexuée. Ce chapitre décrit comment nous avons utilisé l'analyse de concepts formels et ses extensions pour permettre l'exploration de ce réseau. L'ajout de nouvelles informations hétérogènes et l'utilisation de l'analyse de concepts formels nous permet de faire de la fouille de données dans ce réseau afin d'extraire des interactions et des ensembles d'interactions pertinentes pour la problématique étudiée.

Comme il a été montré, le graphe biparti des interactions entre microARN matures et ARNm chez *Acyrthosiphon pisum* peut être considéré comme un contexte formel, appelé par la suite le *contexte des interactions*. L'énumération de l'ensemble des concepts formels sur le contexte des interactions nous permet d'obtenir l'ensemble des microARN matures qui ciblent les mêmes ARNm.

Néanmoins, extraire de la connaissance biologique et interpréter biologiquement ces concepts sans apporter d'informations extérieures, que ce soit sur les interactions ou sur les éléments de ces interactions, se révèle compliqué. Dans l'optique d'associer des informations biologiques aux concepts formels, des attributs biologiques sont rajoutés au contexte entre microARN matures et ARNm. Ces attributs vont nous permettre de caractériser les interactions microARN/ARNm.

La première partie décrit comment des attributs sont ajoutés à un contexte afin de pouvoir interpréter un réseau. Dans une seconde partie cette méthode est appliquée à notre réseau d'interactions chez *A. pisum*.

5.1 Description de l'ajout d'informations hétérogènes à un contexte formel

Pour caractériser les concepts d'un réseau biparti, la méthode nécessite trois contextes formels :

- Le contexte formel qui décrit le réseau biparti par une relation R entre les ensembles d'objets O_1 et O_2 : $\mathbb{K}^{\text{rel}} = (O_1, O_2, R)$;
- Le contexte formel qui décrit les objets O_1 par un ensemble d'attributs A_1 : $\mathbb{K}^1 = (O_1, A_1, I_1)$;
- Le contexte formel qui décrit les objets O_2 par un ensemble d'attributs A_2 : $\mathbb{K}^2 = (O_2, A_2, I_2)$.

À chacun de ces contextes, un ensemble de concepts est associé : $\mathfrak{C}^{\text{rel}}$, $\mathfrak{C}^1$ et $\mathfrak{C}^2$ respectivement pour $\mathbb{K}^{\text{rel}}$, $\mathbb{K}^1$ et $\mathbb{K}^2$.

À partir de ces trois contextes, un nouveau contexte formel $\mathbb{K}^{\text{rel},1,2}$ est formé par la fusion de ces trois contextes sur la base de leurs ensembles communs, O_1 et O_2. Le nouveau contexte est formé de la façon suivante :

$$\mathbb{K}^{\text{rel},1,2} = (O_1 \cup A_2, O_2 \cup A_1, R \cup I_1 \cup \bar{I}_2)$$

avec $\bar{I}_2 \subseteq A_2 \times O_2$ et $(a, b) \in I_2 \Leftrightarrow (b, a) \in \bar{I}_2$.

L'ensemble des concepts $\mathfrak{C}^{\text{rel},1,2}$ de $\mathbb{K}^{\text{rel},1,2}$ à partir des ensembles de concepts $\mathfrak{C}^{\text{rel}}$, $\mathfrak{C}^1$ et $\mathfrak{C}^2$ peut être défini par l'utilisation d'un opérateur adapté à partir de l'opérateur de fusion f défini sur la réparation de contextes bruités (partie 4.1.1). L'ensemble $\mathfrak{C}^{\text{rel}}$ serait considéré comme l'ensemble des concepts d'origine et les ensembles $\mathfrak{C}^1$ et $\mathfrak{C}^2$ comme les « faux » concepts (ce qui n'est pas le cas ici).

Comme l'on souhaite caractériser les relations de $\mathbb{K}^{\text{rel}}$, parmi tous les concepts obtenus sur $\mathbb{K}^{\text{rel},1,2}$, seuls les concepts incluant au moins une relation de R sont intéressants. Ce sont les seuls qui caractérisent potentiellement les interactions du réseau. C'est pourquoi nous définissons un sous ensemble $\mathfrak{CA}^{\text{rel},1,2} \subseteq \mathfrak{C}^{\text{rel},1,2}$ qui distingue l'ensemble des concepts formels de $\mathbb{K}^{\text{rel},1,2}$ qui comprennent une relation d'interaction :

Définition 5.1 *On appelle* concept annoté *un concept du sous ensemble* $\mathfrak{CA}^{rel,1,2} \subseteq \mathfrak{C}^{rel,1,2}$ *défini de la façon suivante :*

$$\forall\, C^{rel,1,2} \in \mathfrak{CA}^{rel,1,2} \Rightarrow \exists o_1 \in extension(C^{rel,1,2}) \wedge \exists o_2 \in intension(C^{rel,1,2}) \wedge (o_1, o_2) \in R,$$

avec $extension(C^{rel,1,2})$ *et* $intension(C^{rel,1,2})$ *l'extension et l'intension de* $C^{rel,1,2}$.

Le sous ensemble de concepts $\mathfrak{CA}^{\text{rel},1,2}$ représente l'ensemble des concepts de $\mathfrak{C}^{\text{rel},1,2}$ qui possèdent au moins une relation de $\mathbb{K}^{\text{rel}}$.

N'obtenir que les concepts $C^{\text{rel},1,2} \in \mathfrak{CA}^{\text{rel},1,2}$ de $\mathbb{K}^{\text{rel},1,2}$ qui possèdent au minimum une relation incluse dans R permet d'obtenir trois types de concepts listés ci-dessous. Dans ce qui suit, $C^{\text{rel},1,2} \in \mathfrak{CA}^{\text{rel},1,2}$, $C^{\text{rel}} = (C, D) \in \mathfrak{C}^{\text{rel}}$, A_1 est l'ensemble des attributs de $\mathbb{K}^1$ et A_2 est l'ensemble des attributs de $\mathbb{K}^2$:

1. $C^{\text{rel},1,2} \in \mathfrak{C}^{\text{rel}}$: le concept ne contient aucune annotation associé sur les ensembles d'attributs A_1 et A_2 ;
2. $C^{\text{rel},1,2} = (D \cup A'_2, E) \wedge A'_2 \subseteq A_2 \wedge A'_2 \neq \emptyset$ ou $C^{\text{rel},1,2} = (D, E \cup A'_1) \wedge A'_1 \subseteq A_1 \wedge A'_1 \neq \emptyset$: le concept peut être associé à un concept de $\mathfrak{C}^{\text{rel}}$ et contient une annotation sur les ensembles d'attributs A_1 ou A_2 ;
3. $C^{\text{rel},1,2} = (D' \cup A'_2, E') \wedge D' \subset D \wedge E' \subset E \wedge D' \neq \emptyset \wedge E' \neq \emptyset \wedge A'_2 \subset A_2 \wedge A'_2 \neq \emptyset$ ou $^{\text{rel},1,2} = (D', E' \cup A'_1) \wedge D' \subset D \wedge E' \subset E \wedge D' \neq \emptyset \wedge E' \neq \emptyset \wedge A'_1 \subset A_1 \wedge A'_1 \neq \emptyset$: le concept contient un rectangle non maximal de $\mathfrak{C}^{\text{rel}}$ et contient une annotation sur les ensembles d'attributs A_1 ou A_2.

Sur le contexte formel $\mathbb{K}^{\text{rel}}$, aucune relation n'est définie sur les ensembles A_2 et A_1, $A_2 \times A_1 = \emptyset$. C'est pour cette raison que les concepts de $\mathfrak{CA}^{\text{rel},1,2}$ incluent soit des attributs de A_2 soit des attributs de A_1 mais pas les deux (cas 2 et 3 ci-dessus). Cela implique qu'il peut exister deux annotations pour un même concept de $\mathfrak{C}^{\text{rel}}$ (cas 2) ou une annotation pour le concept de $\mathfrak{C}^{\text{rel}}$ et une autre sur l'autre ensemble d'attribut un rectangle non maximal de $\mathfrak{C}^{\text{rel}}$ (cas 2,3).

Nous terminerons cette exemple par une illustration. Prenons trois contextes formels : le contexte $\mathbb{K}^{\text{i}}$ sur des interactions microARN/ARNm, le contexte $\mathbb{K}^{\mu}$ sur les attributs des microARN matures et le contexte $\mathbb{K}^{a}$ sur les attributs des ARNm ainsi que le treillis $\underline{\mathfrak{B}}(\mathbb{K}^{\text{i}})$ associé à $\mathbb{K}^{\text{i}}$ (Figure 5.1). Ces trois contextes correspondent aux contextes $\mathbb{K}^{\text{rel}}$, $\mathbb{K}^1$ et $\mathbb{K}^2$ et le treillis représente l'ensemble des concepts $\mathfrak{C}^{\text{rel}}$. La fusion de ces trois contextes donne le contexte $\mathbb{K}^{\text{i},\mu,a}$ et son treillis $\underline{\mathfrak{B}}(\mathbb{K}^{\text{i},\mu,a})$ (Figure 5.2). Ils correspondent au contexte $\mathbb{K}^{\text{rel},1,2}$ et à l'ensemble des concepts $\mathfrak{C}^{\text{rel},1,2}$.

Cet exemple nous permet d'illustrer l'ensemble des cas décrits plus haut. On peut voir que le concept $C_1^{\text{i},1,2}$ en rouge Figure 5.2 ne couvre aucune interaction entre microARN mature et ARNm, il ne fait donc pas partie des concepts d'intérêt pour l'interprétation des interactions microARN/ARNm.

Le cas 1 est illustré par le concept $C_4^{\text{i},1,2}$ qui est déjà présent dans $\mathfrak{C}^{\text{rel}}$ sous le nom C_1^{i}. Il n'est caractérisé par aucun attribut.

Le cas 2 est illustré par le concept $C_7^{\text{i},1,2}$ qui permet de caractériser le microARN du concept $\perp^{\text{i}}$ par l'attribut μ_{att_1}. Le concept C_2^{i} est inclus dans deux concepts, $C_3^{\text{i},1,2}$ et $C_5^{\text{i},1,2}$ où le premier caractérise les ARNm du concept C_2^{i} et le second les microARN de ce concept.

Le cas 3 est illustré notamment par $C_6^{\mathrm{i},1,2}$ qui chevauche le concept C_1^{i}. Il caractérise donc un sous ensemble des relations de C_1^{i} ($\{\mu_1\} \times \{a_1, a_2\}$) par l'attribut μ_{att_2} sur le microARN μ_1.

	a_1	a_2	a_3	a_4
μ_1	1	1		
μ_2	1	1		
μ_3	1	1	1	1
μ_4			1	1
μ_5			1	1

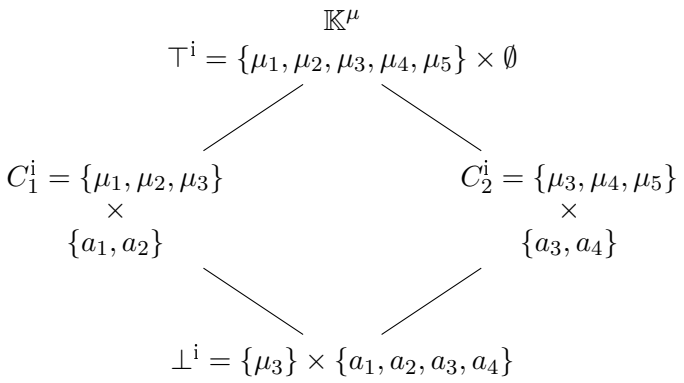
$\mathbb{K}^{\mathrm{i}}$

	μ_{att_1}	μ_{att_2}
μ_1		1
μ_2		
μ_3	1	
μ_4	1	
μ_5	1	1

$\mathbb{K}^{\mu}$

	a_{att_1}	a_{att_2}
a_1		
a_2	1	
a_3		1
a_4		1

$\mathbb{K}^{a}$



$\underline{\mathfrak{B}}(\mathbb{K}_{\mathrm{i}})$

FIG. 5.1 – Trois contextes formels et un treillis de concepts : $\mathbb{K}^{\mathrm{i}}$ représente les interactions microARN/ARNm, $\mathbb{K}^{\mu}$ et $\mathbb{K}^{a}$ représentent respectivement les attributs possédés par les microARN matures $\mu_1..\mu_5$ et les ARNm $a_1..a_4$. Le treillis de concepts $\underline{\mathfrak{B}}(\mathbb{K}_{\mathrm{i}})$ est représenté en dessous.

5.2 Application au réseau d'interactions microARN/ARNm chez *Acyrthosiphon pisum*

L'ajout d'informations biologiques hétérogènes au contexte formel des interactions (matrice d'adjacence du graphe biparti d'interactions microARN/ARNm) par la méthode globalement décrite ci-dessus permet de répondre à deux types de questions :

1. Est-ce que le regroupement d'ensemble de microARN matures et d'ARNm en interaction implique des attributs biologiques spécifiques pour ces deux ensembles ?
2. Est-ce que l'ajout d'attributs biologiques particuliers permet d'observer de nouveaux regroupements d'interactions impliqués par ces attributs ?

Dans la suite de cette partie, les attributs ajoutés au contexte sont tout d'abord détaillés puis on présente une analyse des résultats obtenus par cette méthode sur ces attributs.

	a_1	a_2	a_3	a_4	μ_{att_1}	μ_{att_2}
μ_1	1	1				1
μ_2	1	1				
μ_3	1	1	1	1	1	
μ_4			1	1	1	
μ_5			1	1	1	1
a_{att_1}		1				
a_{att_2}			1	1		

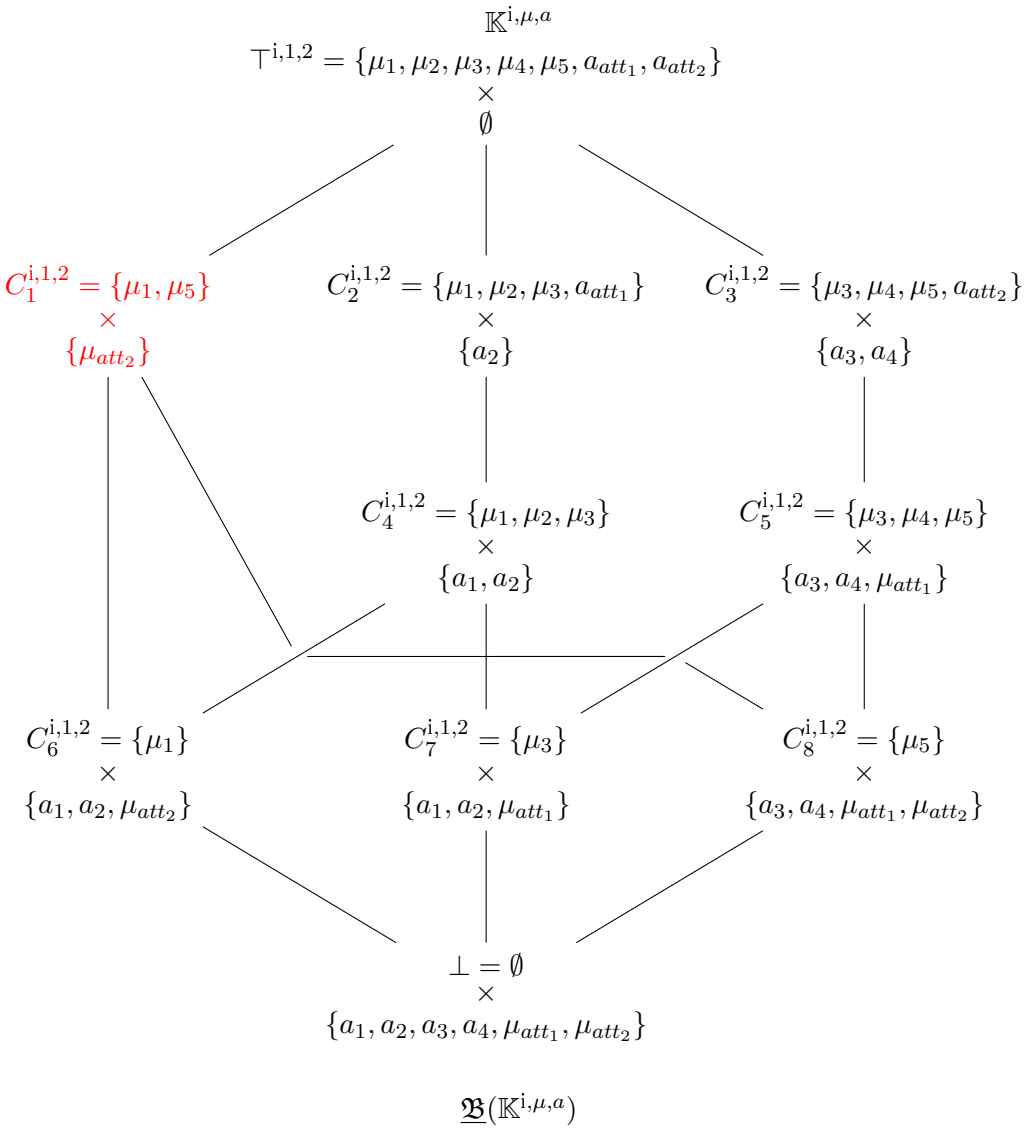


FIG. 5.2 – Le contexte $\mathbb{K}^{i,\mu,a}$ et son treillis de concepts. En rouge le concept qui ne comporte aucune interaction microARN/ARNm et ne fait donc pas partie de $\mathfrak{CA}^{i,\mu,a}$

5.2.1 Attributs biologiques utilisés pour les microARN et les ARNm

Règles de transition de profil cinétique

Cet attribut est identique aux règles issues de la classification des microARN matures et des ARNm différentiellement régulés en fonction de leurs profils cinétiques sexués et asexués décrits chapitre 3. Pour chaque élément du réseau, une règle de transition est associée. Il s'agit d'observer si des microARN matures ou des ARNm possédant des règles de transitions identiques présentent des caractéristiques communes. Chacun des attributs représente l'une des neuf règles de classification et il existe une relation entre l'objet et l'attribut si le couple de profils de l'objet correspond à la règle. Pour les 15 microARN matures il y a exactement un choix effectué parmi les quatre attributs exclusifs associés ce qui génère 15 couples microARN/règle et pour les 1.810 ARNm il y a neuf attributs ce qui génère 1.810 couples microARN/règle.

Éléments différentiellement régulés au premier temps des cinétiques

Il a été montré que passé le stade de développement larvaire 17, les embryons ne sont plus flexibles au kinoprène et que leur avenir, sexué ou asexué, devient déterminé [77]. Le point de transition entre le Stade 17 et les stades suivants mérite une attention particulière. Nous avons donc introduit un attribut pour identifier tout élément régulé possédant une différence de profils d'expression entre le T_0 (Stade 17) et les T_{1S} (sexué) ou les T_{1AS} (asexué). En procédant à cette sélection, huit microARN matures et 475 ARNm ont été marqués par cet attribut intitulé « PCD » (pour Première Cinétique Différente).

Annotation fonctionnelle par GO

Pour chaque ARNm du réseau avec au moins une annotation fonctionnelle GO (pour rappel 690 ARNm), cette ou ces annotations fonctionnelles lui sont associées. Chacune des annotations représente un attribut et il existe une relation entre l'ARNm et l'attribut si l'ARNm est annoté par cette annotation. Il y a 1.409 attributs pour 3.064 relations. Cet attribut est appelé « GO » par la suite.

Annotation fonctionnelle manuelle par termes GO d'intérêt

Une liste de fonctions d'intérêt à observer dans le réseau a été extraite à partir des quatre fonctions biologiques précédemment annotées par Gallot *et al.* [77] et basées sur l'expression différentielle d'ARNm chez *A. pisum* avec un protocole expérimental identique au nôtre. Ces quatre fonctions sont : l'ovogenèse (1), la régulation post-transcriptionnelle (2), l'épigénétique (3) et le cycle cellulaire (4). À ces fonctions sont ajoutées trois autres fonctions : le système neuroendocrine (5), le développement musculaire (6) et la régulation transcriptionnelle (7).

Pour la fonction 5, il a été décidé de l'ajouter car ces éléments jouent un rôle avéré ou potentiel dans la réception et transduction du signal de la photopériode chez le puceron du pois [73]. La fonction 6 a été ajoutée car l'on sait que du point de vue phénotypique les femelles sexuées et asexuées ne sont pas identiques, d'où une différence de régulation potentielle des ARNm impliqués dans le développement musculaire. La fonction 7 a été

ajoutée car on a constaté un enrichissement en termes GO associés à cette fonction dans les ARNm différentiellement régulés (pour rappel, voir Tableau 3.3 partie 3.2).

Une fois ces sept fonctions définies, elles ont été caractérisées par des ensembles de termes GO. Si un ARNm possède un des termes GO associés à l'une des fonctions, alors il est aussi associé à cette fonction. L'ensemble des sept fonctions et leurs annotations et numéro GO sont listées ci-dessous :

1. Ovogenèse : oogenesis (0048477), gonad development (0008406), reproductive structure development (0048608) ;
2. Régulation post-transcriptionnelle : posttranscriptional regulation of gene expression (0010608), production of small RNA involved in gene silencing by RNA (0070918) ;
3. Épigénétique : regulation of molecular function, epigenetic (0040030), regulation of gene expression, epigenetic (0040029), DNA methylation (0006306), chromatin organization (0006325), chromatin silencing (0006342) ;
4. Cycle cellulaire : cell cycle (0007049), cell differentiation (0030154), cell cycle DNA replication (0044786) ;
5. Système neuroendocrine : nervous system development (0007399), synaptic transmission (0007268), response to hormone (0009725), regulation of hormone levels (0010817), neuropeptide signaling pathway (0007218), neuropeptide catabolic process (0010813), neurotransmitter transport (0006836), dopamine secretion (0014046), dopamine metabolic process (0042417), octopamine secretion (0061539), octopamine metabolic process (0046333), histamine metabolic process (0001692), response to histamine (0034776), histamine transport (0051608), ecdysone metabolic process (0008205), response to ecdysone (0035075), mevalonate transport (0015728), isoprenoid biosynthetic process via mevalonate (1902767), terpenoid biosynthetic process, mevalonate-dependent (0051485), steroid metabolic process (0008202), steroid hormone secretion (0035929) ;
6. Développement musculaire : muscle structure development (0061061) ;
7. Régulation transcriptionnelle : regulation of transcription, DNA-templated (0006355).

Le Tableau 5.1 donne le nombre d'ARNm associés à chacune des fonctions sur les ARNm qui possèdent des cinétiques différentes. À noter que certains ARNm peuvent être associés à plusieurs fonctions, c'est pourquoi le nombre total d'ARNm est différent de la somme des ARNm associés à chacune des fonctions.

fonction	nombre d'ARNm
ovogenèse	20
régulation post-transcriptionnelle	7
épigénétique	19
cycle cellulaire	97
système neuroendocrine	67
développement musculaire	12
régulation transcriptionnelle	52
total	274
ARNm unique associé à une fonction	146

Tableau 5.1 – Nombre d'ARNm différentiellement régulés impliqués dans chaque fonction.

Chacune de ces fonctions représente un attribut et il existe une relation entre l'ARNm et l'attribut si l'ARNm est annoté par cette fonction. Il y a 7 attributs pour 274 relations. Dans la suite de la thèse cet attribut est appelé « fonction manuelle ».

Aucun attribut de fonction n'a été ajouté pour les microARN matures à cause de leur faible effectif.

5.2.2 Résultat sur le réseau d'interaction microARN/ARNm chez *Acyrthosiphon pisum*

La méthode d'enrichissement des modules d'interaction a été appliquée au réseau sur les contextes suivants :

- Le contexte des interactions entre microARN matures et ARNm : 15 microARN matures, 1.810 ARNm et 2.250 interactions ;
- Le contexte des attributs des microARN matures : 15 microARN matures, 5 attributs et 23 couples en relations ;
- Le contexte des attributs des ARNm : 1.810 ARNm, 1.426 attributs et 5.623 couples en relations.

Le nombre de concepts, ou modules, obtenus contenant au moins une interaction microARN/ARNm est de 2.225. Une première analyse rapide de ces concepts montre que l'introduction de l'attribut « GO » rend le nombre de concepts vraiment trop important (2.225) et tend à brouiller l'interprétation du réseau. Ceci peut être dû au fait qu'il n'y a que 690 ARNm sur les 1.810 qui possèdent au moins un attribut GO et que la précision des annotations n'est pas la même pour ces 690 ARNm. Pour cette raison, il a été décidé de supprimer l'attribut portant sur l'annotation fonctionnelle par GO, ce qui ramène le contexte des attributs des ARNm à : 1.810 ARNm, 17 attributs et 2.559 relations.

Le nombre de concepts obtenus en supprimant l'annotation GO de l'analyse passe à 555 concepts, ce qui est très inférieur au nombre précédent (2.225). Sur l'ensemble des concepts obtenus en excluant le top (supremum du treillis) et le bottom (infimum du treillis), 65 comportent au moins un attribut sur les microARN matures, 473 au moins un attribut sur les ARNm et 15 ne comportent aucun attribut.

Avant d'aborder l'analyse proprement dite, nous présentons Figure 5.3 un exemple du type de résultat que l'on peut obtenir. Elle est issue de l'extraction de cinq concepts du treillis. Chaque concept est constitué de cinq parties (séparées par un trait) toujours présentées dans cet ordre :

1. Le numéro du concept ;
2. L'ensemble des attributs possédés par les ARNm ;
3. L'ensemble des ARNm (nomenclature AphidBase : ACYPI). Si le nombre d'ARNm présents dans le concept dépasse les 25, alors dans ce cas seul le nombre d'ARNm présents est affiché ;
4. L'ensemble des microARN matures ;
5. L'ensemble des attributs possédés par les microARN matures.

À noter que pour faciliter la lisibilité et l'interprétation des figures, certains attributs ont été ajoutés aux concepts après la création du treillis. Ces cas correspondent aux

concepts possédant une annotation à la fois sur les microARN matures et les ARNm (cas 2 et cas 2,3). Les attributs sont donc ajoutés si et seulement si ils sont présents pour l'ensemble des éléments (microARN matures ou ARNm du concept). Les attributs ajoutés apparaissent entre accolades.

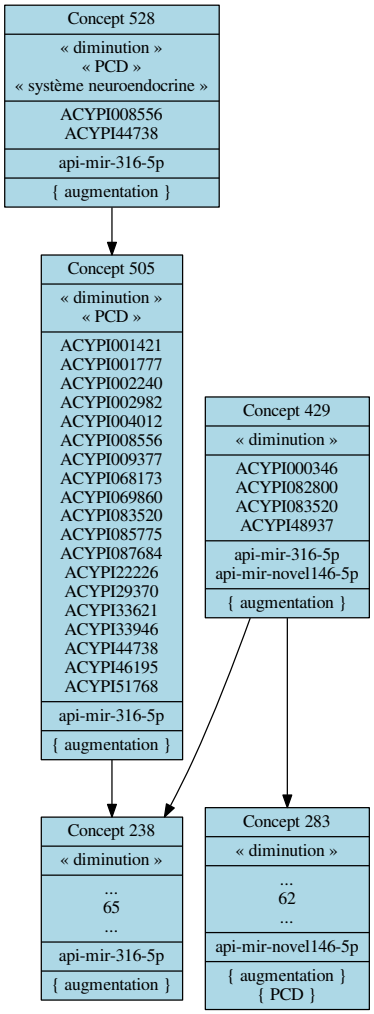


FIG. 5.3 – Cinq concepts extraits du treillis des interactions après ajout d'information.

Une des façons de lire cette figure est la suivante : le concept 238 inclut 65 ARNm ciblés par le microARN api-mir-316-5p, l'ensemble des ARNm suivent la règle « diminution » et api-mir-316-5p suit la règle « augmentation ». Ensuite, le concept 283 regroupe 62 ARNm qui suivent la règle « diminution » et sont ciblés par api-mir-novel146-5p qui lui possède les attributs « augmentation » et « PCD » (élément différentiellement régulé au premier pas de temps des cinétiques). Le concept 429 regroupe quatre cibles communes de api-mir-316-5p et api-mir-novel146-5p qui suivent la règle « diminution ». On peut voir aussi que ces deux microARN suivent la règle « augmentation ». Les quatre ARNm du concept 429 forment un sous ensemble à la fois des ARNm du concept 238 et des ARNm du concept 283. Le concept 505 regroupe un sous ensemble des ARNm ciblés par api-mir-316-5p et présents dans le concept 238 mais où cette fois-ci l'ensemble

de ces ARNm possèdent les attributs « diminution » et « PCD ». De la même façon, les ARNm du concept 528 (sous ensemble de ceux présents dans le concept 505) possèdent toujours les attributs « diminution » et « PCD » mais aussi l'attribut « système neuroendocrine ». Les figures ont été produites à l'aide de graphviz [162].

Exploration non supervisée des modules

L'exploration non supervisée consiste à observer l'ensemble des modules formés sans *a priori* biologique en regardant préférentiellement des modules respectant certaines contraintes. On peut ainsi chercher à observer les modules contenant au minimum deux microARN matures et deux ARNm ou encore contraindre que les règles suivies par les microARN matures soit cohérentes à celles suivies par les ARNm.

Exploration des modules contenant au minimum deux microARN matures et deux ARNm Dans cette sous-partie nous nous intéressons aux concepts contenant au moins deux microARN matures et au moins deux ARNm pour observer si certains microARN matures ciblent des ensembles d'ARNm identiques qui seraient impliqués dans des fonctions identiques.

Sur l'ensemble des 555 concepts obtenus, un tiers (164) possèdent au minimum deux microARN matures et deux ARNm. Les sous-graphes comportant uniquement ces concepts ont été extraits. Ils se divisent en 20 composantes connexes, où une composante connexe regroupe tous les concepts reliés par un chemin du sous-graphe. Autrement dit, il n'existe aucune relation entre les concepts de deux composantes connexes différentes. Les composantes connexes sont majoritairement composées d'un seul concept (12 composantes), six composantes incluent deux concepts, une composante inclut trois concepts et la dernière composante inclut 137 concepts. En observant cette dernière composante, il apparaît qu'elle est composée quasi exclusivement de concepts incluant le microARN api-mir-3019-5p, qui pour rappel cible 54 % des ARNm du réseau. Interpréter la fonction de ce microARN mature semble difficile au vu de son nombre énorme d'interactions. Des critères sur les modules précédemment extraits ont donc été rajoutés pour déterminer si un concept devait être gardé ou non pour cette analyse :

- Un module qui ne contient que deux microARN ne doit pas contenir api-mir-3019-5p ;
- Un module qui contient strictement plus de deux microARN peut contenir api-mir-3019-5p.

Une fois ces critères ajoutés, le nombre de concepts, le nombre de composantes et la répartition des concepts au sein des composantes ne comporte plus que 76 concepts répartis dans 29 composantes connexes avec :

- 14 concepts seuls ;
- quatre composantes avec deux concepts ;
- cinq composantes avec quatre concepts ;
- deux composantes avec cinq concepts ;
- quatre composantes avec six concepts.

On note tout d'abord que le nombre maximum de microARN matures dans un module est de trois et que les microARN api-mir-281-5p, api-mir-278-5p, api-mir-3026-5p et api-mir-novel85-3p ne sont pas présents, ce qui fait que seuls 11 microARN matures

apparaissent dans les concepts respectant les critères définis ci-dessus.

Sur l'ensemble de ces concepts, une première analyse possible est d'observer la cooccurrence des différents microARN matures dans ces concepts : quels sont les microARN matures qui se retrouvent au sein des mêmes concepts ? En d'autres termes, est-ce qu'il existe des microARN matures qui ont tendance à se regrouper du fait de leurs interactions ? Parmi les microARN qui apparaissent souvent ensemble (sans prendre en compte api-mir-3019-5p) nous pouvons citer api-mir-1000-5p et api-mir-novel146-5p, api-mir-1-3p et api-mir-1000-5p, api-mir-14-3p et api-mir-316-5p, api-mir-316-5p et api-mir-novel146-5p. On peut noter la cooccurrence de api-mir-1000-5p et api-mir-263a-5p, deux microARN matures exprimés dans la tête soit chez l'abeille [137] soit chez le ver à soie [56] (Figure 5.4). Dans ce sous-graphe, api-mir-1000-5p et api-mir-263a-5p ciblent trois ARNm communs et l'ajout de api-mir-3019-5p réduit ce nombre à deux. On peut observer que les concepts 36 et 37 peuvent être regroupés, de même pour les concepts 34 et 35. Les trois ARNm associés suivent la règle « augmentation » et les deux microARN la règle « diminution » et possèdent aussi l'attribut « PCD ». Si l'on regarde de plus près les annotations des ARNm, on peut voir que ACYPI000235 est annoté comme « protein held out wings-like », ACYPI005514 comme « beta-amyloid-like protein » (attribut fonction : « cycle cellulaire » et « système neuroendocrine ») et que ACYPI002763 ne possède aucune annotation. En regardant plus en détails les annotations GO de ACYPI005514, on voit que ce gène est impliqué dans le développement du système nerveux périphérique, ce qui est cohérent avec la localisation de api-mir-1000-5p et api-mir-263a-5p dans la tête. Les sites de fixation de ces microARN matures sur ces ARNm sont trop distants, entre 150 et 450 nucléotides, pour pouvoir faire l'hypothèse d'une coopération entre ces sites. En plus de cette cooccurrence, celle de api-mir-14-3p et api-mir-263a-5p, deux microARN matures potentiellement impliqués dans l'apoptose, se retrouvent de façon intéressante au sein du même concept (le concept 456 Figure 5.5). Ce concept inclut deux ARNm : ACYPI004009 (« augmentation ») et ACYPI008827 (« diminution »), tous deux annotés respectivement comme une « protein phosphatase 1 regulatory subunit 37 » et une « forkhead box protein o-like » (facteur de transcription). On peut noter que les protéines de la famille FOXO3, une sous-famille des protéines forkhead box, sont impliquées dans l'apoptose [163], ce qui était déjà le cas pour les deux microARN matures. L'hypothèse d'une action coopérative des sites de fixation sur l'un des deux ou les deux ARNm est tentante, mais les sites semblent trop éloignés pour cela (183 nucléotides pour ACYPI008827 et 96 pour ACYPI004009).

Hormis ces deux cas, il ne semble pas y avoir d'autres microARN matures partageant des fonctions/annotations connues semblables qui se regroupent préférentiellement. Une autre information à noter est que parmi les 76 concepts, aucun ensemble d'ARNm ne partage la même fonction parmi celles manuellement sélectionnées.

Exploration des modules contenant des interactions aux règles cohérentes
Suivant la règle suivie par un microARN mature, on peut définir un ensemble de règles dites cohérentes qui, si ces règles sont suivies par des cibles potentielles de ce microARN mature, sont cohérentes avec une régulation de l'expression de ces cibles par le microARN. Par exemple, si un microARN mature suit la règle « diminution », alors on s'attend à ce que les ARNm qu'il régule suivent des règles comme « augmentation », « avance » ou encore « disparition pic négatif ». Le Tableau 5.2 présente l'ensemble des

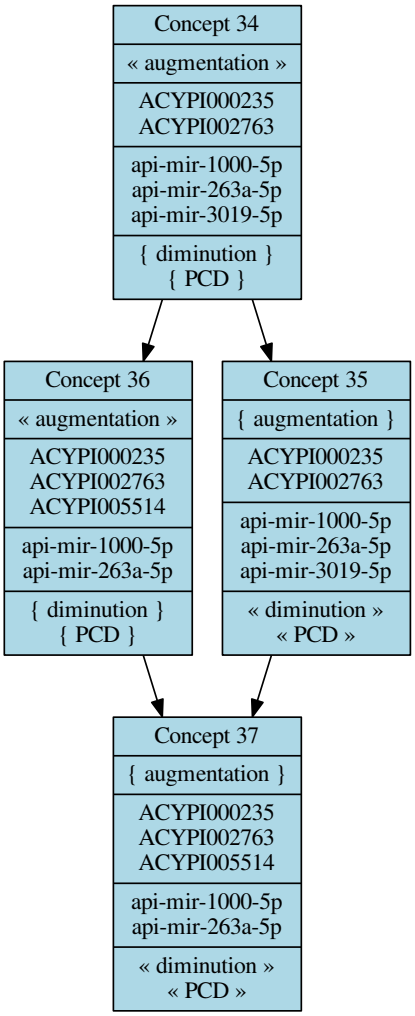


FIG. 5.4 – Quatre concepts avec la cooccurrence de api-mir-1000-5p et api-mir-263a-5p.

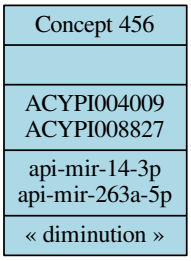


FIG. 5.5 – Concept avec la cooccurrence de api-mir-14-5p et api-mir-263a-5p.

règles dites cohérentes.

Sur l'ensemble des 555 concepts, 161 (29 %) comportent uniquement des microARN matures et des ARNm avec des règles cohérentes et les sous-graphes comportant uniquement ces concepts ont été extraits. Ces sous-graphes se divisent en 12 composantes connexes. Les 161 concepts se répartissent de la façon suivante dans les 12 compo-

microARN	ARNm
diminution	augmentation avance disparition pic négatif apparition pic positif
augmentation	diminution retard apparition pic négatif disparition pic positif
retard	retard

Tableau 5.2 – Listes des règles cohérentes.

santes connexes : huit composantes incluant un concept, une composante incluant deux concepts, une composante incluant quatre concepts, une composante incluant 10 concepts et une composante incluant 137 concepts.

La Figure 5.6 présente les huit composantes connexes ne contenant qu'un seul concept. Les microARN matures représentés dans ces concepts sont au nombre de cinq : api-mir-263a-5p, api-mir-3019-5p, api-mir-1000-5p, api-mir-316-5p et api-mir-34-5p, chacun apparaissant dans un ou deux concepts. Ces concepts ne sont constitués que d'un seul microARN mature et aucune fonction n'est partagée par l'ensemble des ARNm d'un de ces modules ce qui rend l'interprétation de ces modules difficile. On peut néanmoins noter dans le concept 310 (en haut à gauche Figure 5.6) la présence de l'ARNm ACYPI005313 annoté comme codant pour une protéine de la famille des facteurs de croissance : la protéine TGF-β1, qui a été montrée comme étant notamment impliquée dans l'apoptose chez la souris [164]. Le microARN mature appartenant à ce concept est api-mir-263a-5p qui a été identifié comme étant lui aussi impliqué dans l'apoptose chez la drosophile [17]. Le deuxième ARNm du module, ACYPI008685, est lui annoté comme une phosphorylase b kinase, ce qui ne semble pas avoir de rapport avec l'apoptose. Ce microARN fait aussi partie du concept 384 où l'ARNm ACYPI003493 est annoté comme codant pour une « snf-related serine threonine-protein kinase » (SNRK), protéine qui pourrait être impliquée elle aussi dans l'apoptose chez le rat [165]. Concernant les autres concepts, aucune information sur les ARNm ou sur les microARN matures ne semble pertinente pour une analyse plus poussée.

La Figure 5.7 présente les composantes connexes avec deux concepts (à gauche) et quatre concepts (à droite). Chaque sous-graphe implique encore une fois uniquement un microARN mature, api-mir-3038-3p, spécifique au puceron et api-mir-316-5p sur lequel aucune information fonctionnelle n'est disponible. Ces deux sous-graphes ne nous apportent pas beaucoup d'information pour venir enrichir notre connaissance sauf en ce qui concerne les deux concepts 230 et 388 où l'attribut « développement musculaire » est présent et où les attributs « cycle cellulaire » et « ovogénèse » sont présents dans le concept 230.

La Figure 5.8 présente la composante connexe avec 10 concepts. Cette fois-ci deux microARN matures sont présents dans ce sous-graphe : api-mir-316-5p et api-mir-novel146-5p. On peut le diviser en trois parties : la partie de gauche impliquant uniquement api-mir-316-5p (cadre api-mir-316-5p), la partie de droite impliquant uniquement

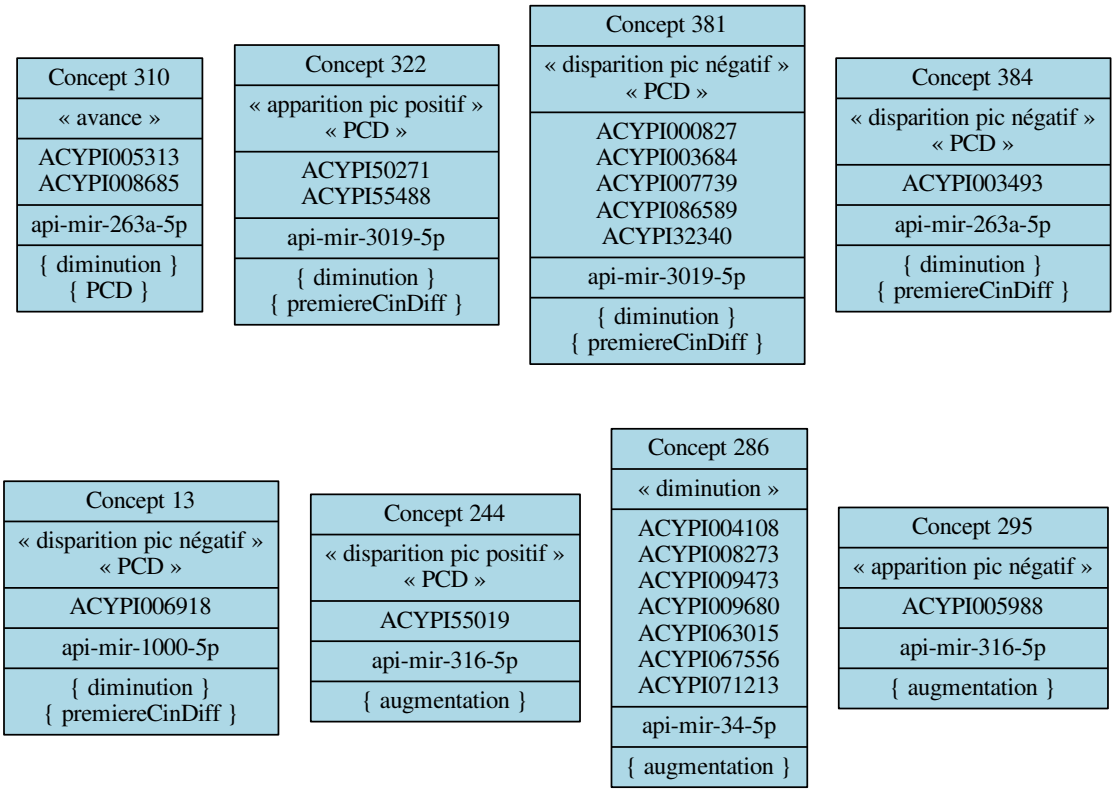


FIG. 5.6 – Les huit composantes connexes avec un seul concept issues de la recherche de modules contenant uniquement des règles cohérentes.

api-mir-novel146-5p (cadre api-mir-novel146-5p) et le concept 429 au centre avec les deux microARN mature connectant les deux blocs. On peut noter la présence dans le concept 429 de l'ARNm ACYPI000346, une « GTP-binding protein RHES-like » où la protéine RHES semble impliquée dans la neurotransmission de la dopamine chez la souris [166], fonction qui est liée au système neuroendocrine présent dans les concepts 236, 197, 528, 199 et 200.

Pour la dernière composante, celle incluant 137 concepts, le nombre de concepts ne permet pas de générer une image imprimable dans le document. De plus, aucun module ne ressort de la composante par son nombre d'interactions ou par un aspect particulier.

Exploration supervisée des modules

Après une exploration « sans *a priori* », l'exploration supervisée des modules consiste cette fois-ci à utiliser des attributs biologiques, comme ici les fonctions manuelles ou encore l'attribut PCD.

Exploration des modules contenant des fonctions définies manuellement Pour chacune des fonctions biologiques définies manuellement, les concepts qui possèdent cette fonction ont été extraits ainsi que les composantes connexes impliquant cette fonction. À la différence des extractions précédentes, les concepts directement inférieurs aux

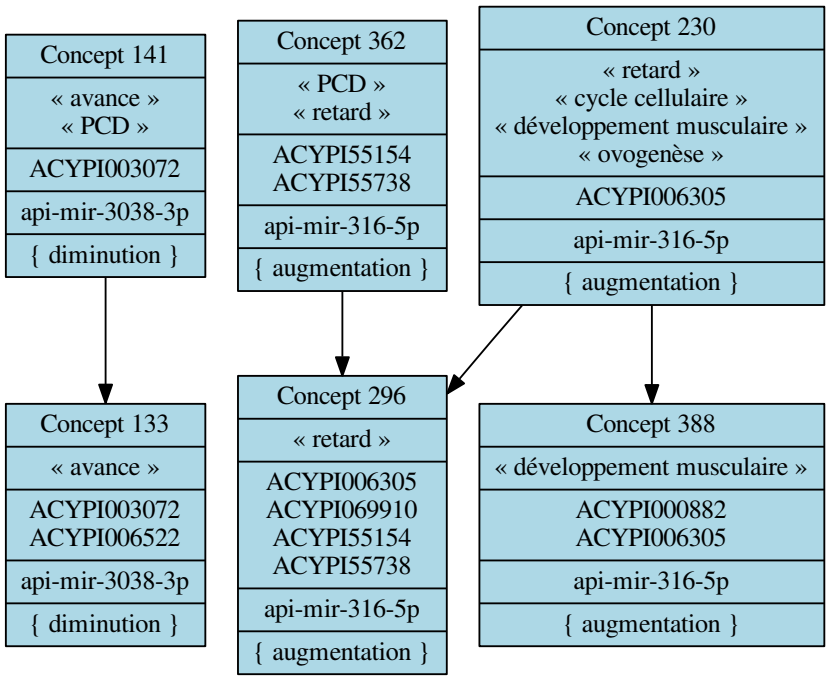


FIG. 5.7 – Les deux composantes connexes avec deux et quatre concepts issues de le recherche de modules contenant uniquement des règles cohérentes.

concepts contenant la fonction ont aussi été extraits. Par exemple dans la Figure 5.9 qui concerne des modules liés à l'ovogenèse, le concept 77 ne possède pas cet attribut mais il est directement relié à l'un d'eux. Cette sélection étendue tient compte du fait que certains ARNm peuvent ne pas être annotés par une fonction précise (pas d'annotation GO couvrant cette fonction ou annotations incomplètes) mais participe quand même à ce processus biologique. Pour l'ensemble des sept fonctions, un résumé des résultats sur le nombre de concepts et le nombre de composantes connexes est présenté Tableau 5.3.

fonction	nombre de concepts	nombre de composantes connexes	nombre d'ARNm
ovogenèse	61	3	20
régulation post-transcriptionnelle	26	2	7
épigénétique	86	1	19
cycle cellulaire	239	1	97
système neuroendocrine	208	2	67
développement musculaire	48	3	12
régulation transcriptionnelle	171	3	52

Tableau 5.3 – Nombre de concepts couvrant l'une des sept fonctions manuelles et nombre de composantes connexes extraites du treillis associé à ces concepts.

Ce tableau donne à lui seul une vision globale des grandes fonctions biologiques représentées dans le réseau d'interactions. Ceci signifie que nos travaux ont cerné des

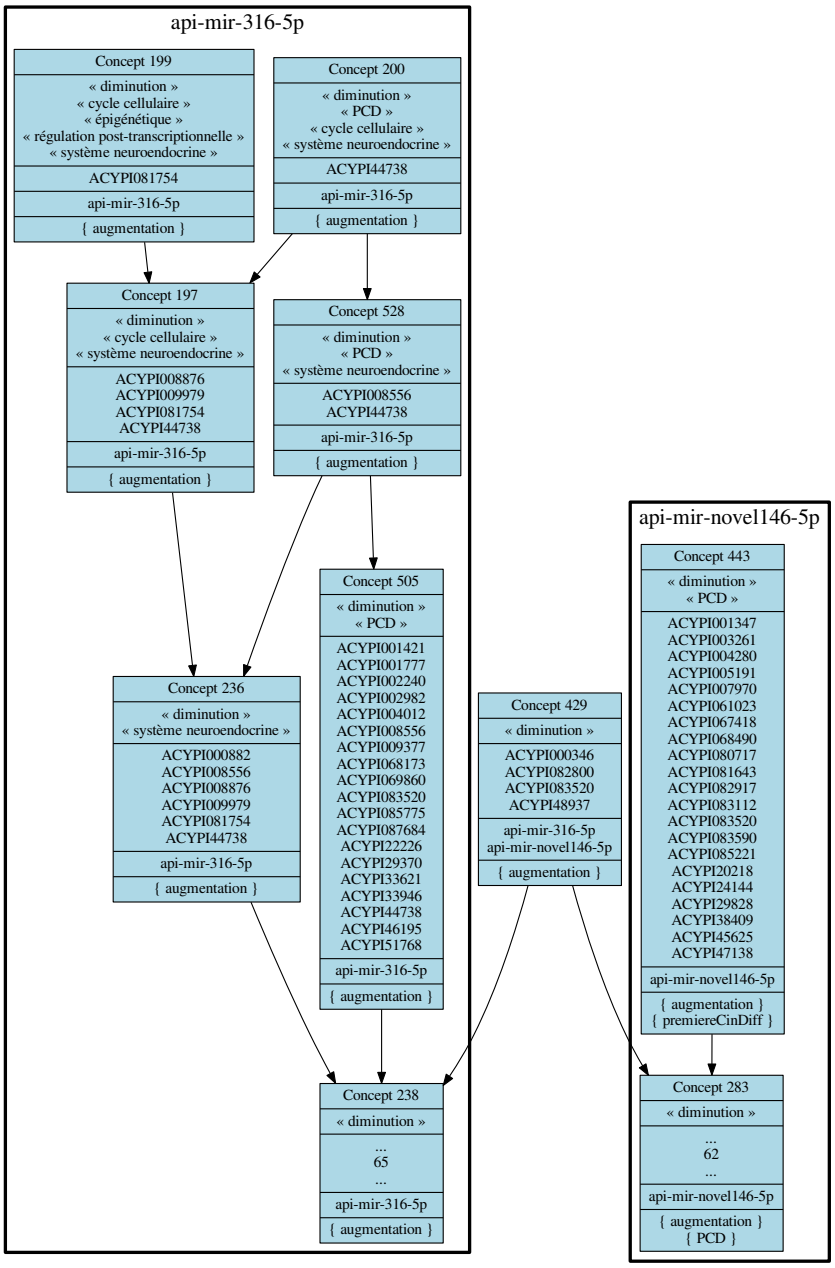


FIG. 5.8 – La composante connexe avec 10 concepts issue de le recherche de modules contenant uniquement des règles cohérentes. Les concepts encadrés sont ceux qui n'impliquent qu'un seul microARN mature.

fonctions potentiellement importantes dans la discrimination des embryogenèses sexuées et asexuées chez le puceron du pois lors de la plasticité phénotypique du mode de reproduction. Les 4 premières fonctions avaient déjà été identifiées lors d'un travail précédent [77] et correspondent à des fonctions actives lors de l'ovogenèse précoce, dans laquelle des régulations épigénétiques et post-transcriptionnelles ont lieu. Les embryons étudiés développent soit des gamètes méiotiques vraies (embryogenèse sexuée) soit des gamètes

diploïdes non méiotiques (embryogenèse asexuée) : il est donc cohérent que la fonction cycle cellulaire soit représentée. Les embryons en formation finissent de construire les tissus et la présence du développement musculaire et du système nerveux n'est donc pas inattendue. De plus, il est connu que les régulations neuroendocrines jouent un rôle important dans la plasticité phénotypique. Enfin, la forte signature sur la régulation transcriptionnelle souligne encore une fois un mécanisme biologique développemental dynamique dans lequel de nombreux programmes génétiques sont régulés. D'ailleurs, la présence de nombreux facteurs de transcription est à souligner, sachant que ces protéines forment souvent des boucles de régulations tripartites entre microARN, ARNm et facteurs de transcription [167].

On peut voir que le nombre de concepts varie grandement entre les différentes fonctions, entre 26 pour la régulation post-transcriptionnelle et 239 pour le cycle cellulaire. Cette taille suit le nombre d'ARNm annotés par ces fonctions, ce qui est cohérent. Par contre le nombre de composantes connexes, entre une et trois, ne suit ni le nombre de concepts ni le nombre d'ARNm. Pour toutes les fonctions où il y a plus d'une composante connexe, il y a toujours une grande composante impliquant à la fois la majorité des concepts et api-mir-3019-5p et d'autres composantes plus petites, d'une taille comprise entre 2 et 19 concepts. Il a été choisi d'exposer plus en détails certains sous-graphes et concepts issus de ceux couvrant les attributs « ovogénèse » et « régulation post-transcriptionnelle » car ces sous-graphes sont facilement interprétables et comportent des modules intéressants.

La fonction « ovogénèse » est composée de trois composantes connexes contenant respectivement 4, 19 et 38 concepts. Ces trois composantes couvrent les microARN matures api-mir-316-5p pour la première, api-mir-1000-5p et api-mir-3038-3p pour la deuxième et api-mir-3019-5p, api-mir-novel146-5p et api-mir-263a-5p pour la troisième. La première composante ne présente pas de caractère particulier, de même que la troisième composante qui est très grande et qui met en jeu quasi exclusivement api-mir-3019-5p. La deuxième composante, présentée Figure 5.9, 5.10 et 5.11, possède des caractéristiques intéressantes. Cette composante implique principalement api-mir-1000-5p et api-mir-3038-3p. Néanmoins, api-mir-3038-3p n'est présent que dans trois concepts : 8, 55 et 134 (Figure 5.10). Les concepts 77 et 93 (Figure 5.11) ne possèdent pas l'attribut « ovogénèse » mais les attributs « cycle cellulaire » et « régulation transcriptionnelle » pour 77 et « épigénétique » pour 93. De plus, ces deux concepts sont les deux concepts les plus bas dans la partie du sous-graphe impliquant uniquement api-mir-1000-5p, et l'ensemble des autres concepts impliquant api-mir-1000-5p sont reliés à au moins l'un de ces deux concepts. Ce qui signifie que, avec les données d'annotation disponibles, tous les ARNm ciblés par api-mir-1000-5p et qui sont connus pour être impliqués dans l'ovogénèse sont aussi impliqués soit dans le cycle cellulaire et la régulation transcriptionnelle soit dans le contrôle épigénétique.

La fonction « régulation post-transcriptionnelle » est pour sa part constituée d'une composante de 3 concepts et d'une deuxième de 23 concepts. La première composante n'implique que api-mir-316-5p et seulement un concept implique cette fonction. Ce concept est composé d'un seul ARNm, ACYPI081754, qui est annoté comme « nuclear cap-binding protein subunit 1-like » et possède des annotations GO en lien avec le coiffage des ARNm, la production de petits ARN interférents ou encore l'inhibition de gènes par les microARN. La deuxième composante, présentée Figure 5.12, 5.13 et 5.14,

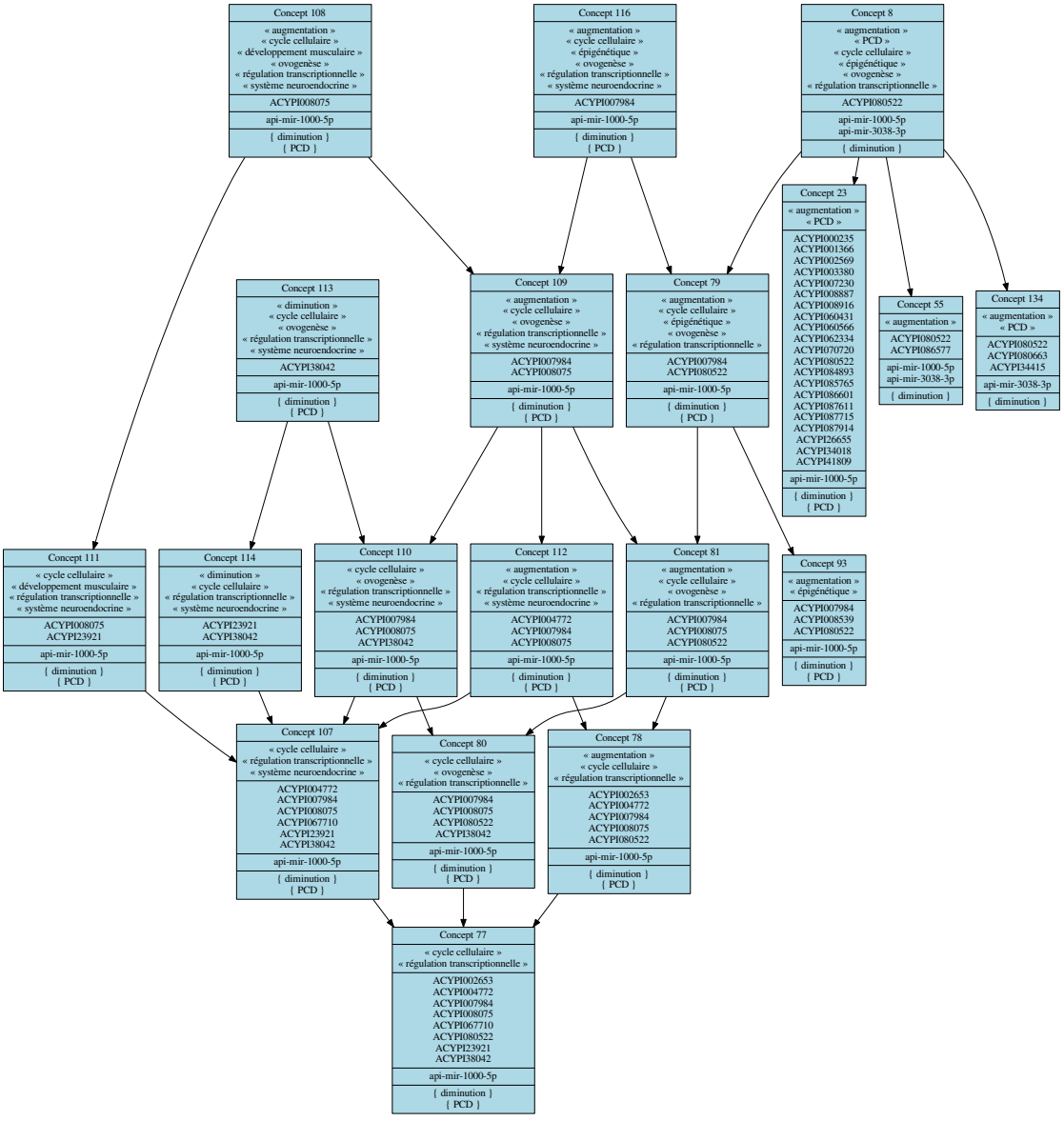


FIG. 5.9 – Deuxième composante connexe impliquant la fonction « ovogénèse ».

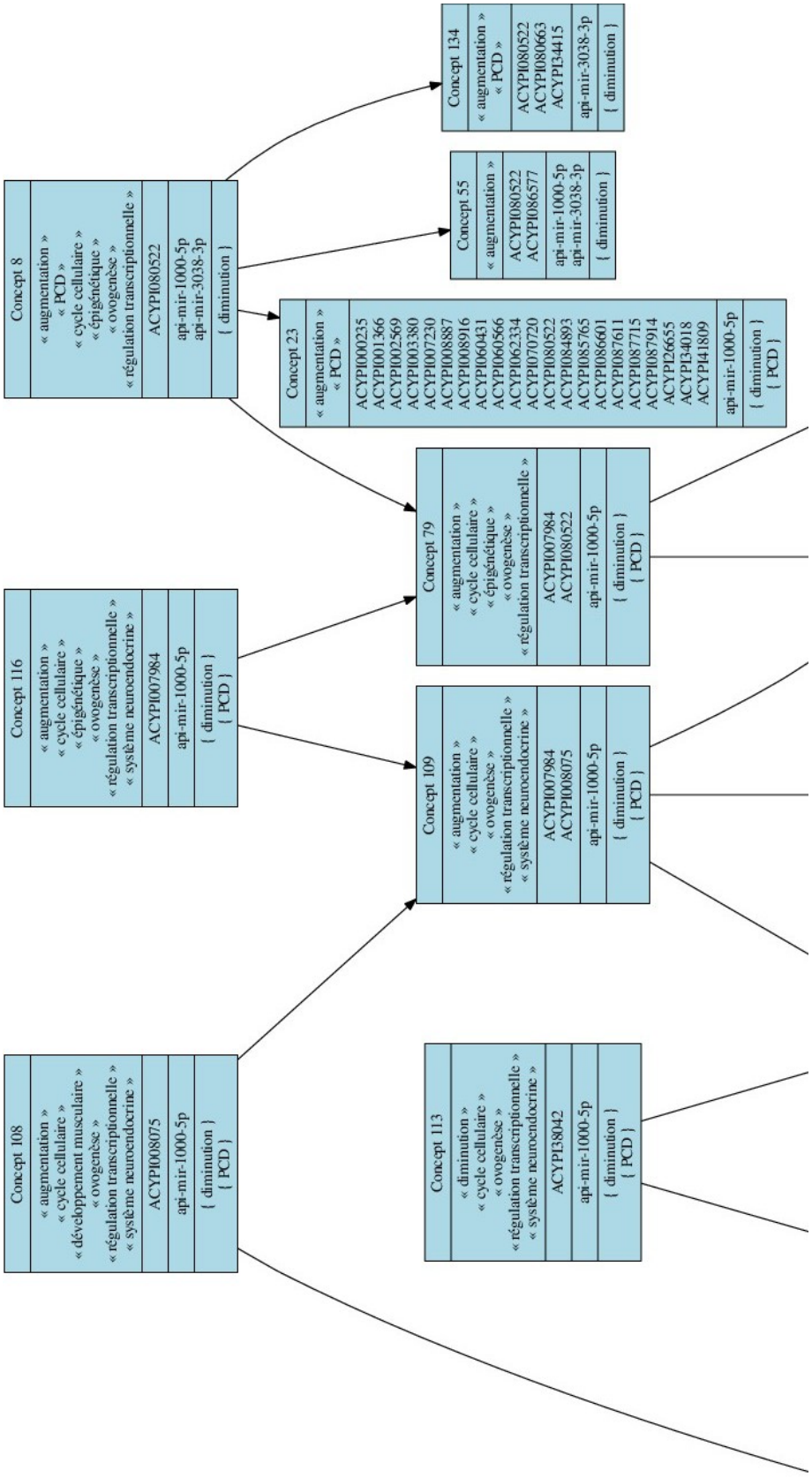


FIG. 5.10 – Zoom de la partie haute de la deuxième composante connexe impliquant la fonction « ovogénèse ». Correspond aux deux premiers niveaux.

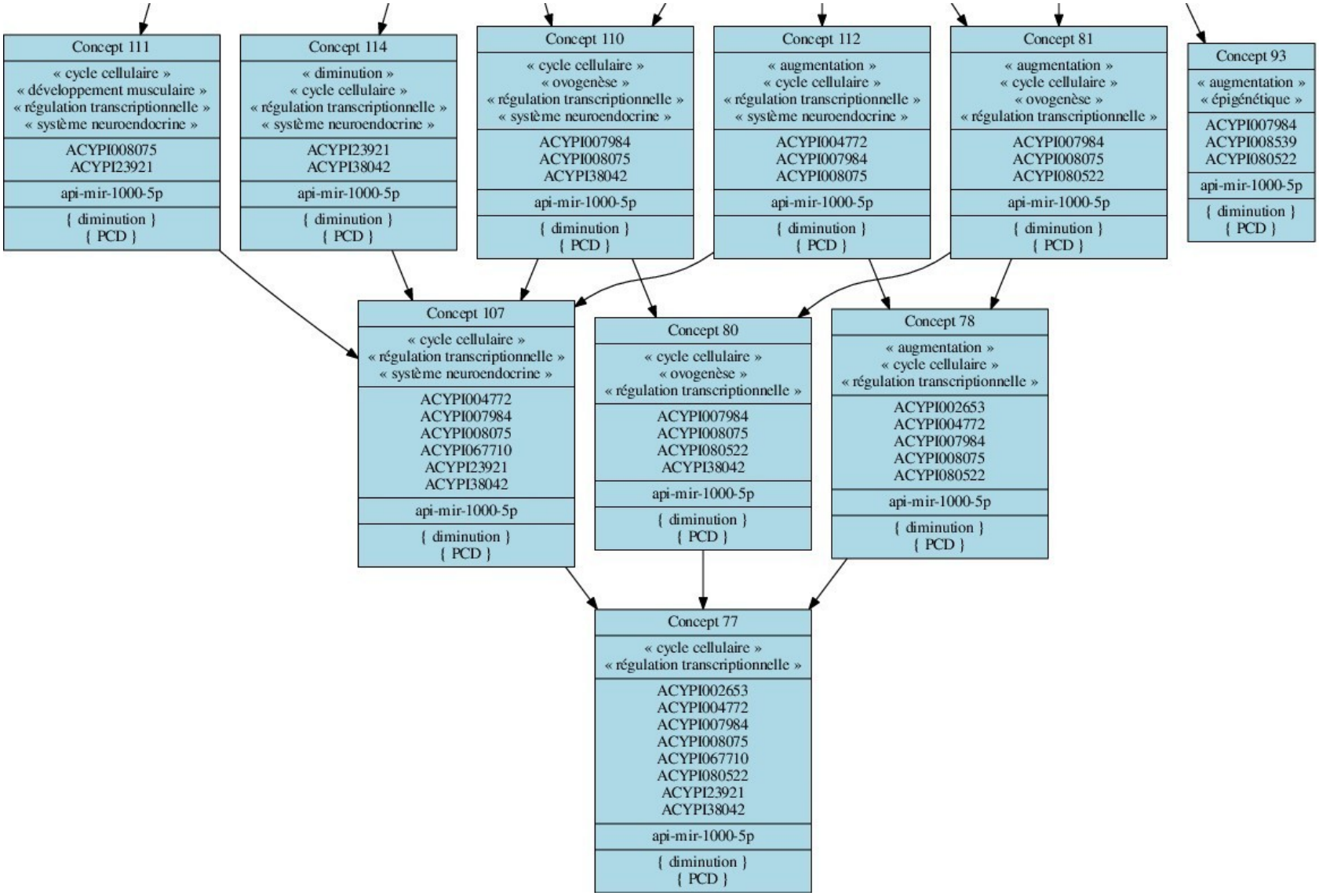


FIG. 5.11 – Zoom de la partie basse de la deuxième composante connexe impliquant la fonction « ovogénèse ». Correspond aux trois derniers niveaux.

implique api-mir-1-3p, api-mir-1000-5p et api-mir-3019-5p. Les deux concepts 41 et 554 les plus bas dans le sous-graphe impliquant api-mir-1000-5p et api-mir-3019-5p incluent l'attribut « augmentation », ce qui signifie que l'ensemble des ARNm ciblés par ces deux microARN matures suivent cette règle, qui est cohérente avec celle des deux microARN matures qui possèdent la règle « diminution ». Deux concepts, 117 et 82 (Figure 5.13), impliquent chacun deux microARN matures respectivement api-mir-1-3p, api-mir-1000-5p et api-mir-1000-5p et api-mir-3019-5p. Le concept 117 inclut l'ARNm ACYPI004772 annoté comme « ribosomal protein s6 kinase alpha-3-like » et impliqué dans la transduction du signal. Il pourrait réguler l'activité de plusieurs facteurs de transcription. Les sites de fixation sont trop éloignés l'un de l'autre (161 nucléotides) pour permettre une coopération entre eux. Le concept 82 inclut l'ARNm ACYPI008539 annoté comme « arsenite-resistance » aussi appelé « serrate RNA effector molecule homolog ». Il est montré comme étant impliqué positivement dans la biosynthèse des microARN chez *Drosophila* et la souris [168, 169]. Deux sites de fixation sont prédits pour api-mir-3019-5p et un pour api-mir-1000-5p, qui est à une distance de 15 nucléotides de l'un des sites de api-mir-3019-5p. Cette distance est compatible avec une coopération entre sites de fixation. On peut donc supposer une action conjointe de ces deux microARN matures sur ACYPI008539.

Exploration des modules contenant des éléments régulés différentiellement lors du premier temps des cinétiques L'attribut « PCD » correspond à la transition entre le stade embryonnaire encore sensible aux conditions changeantes du milieu et les stades ultérieurs du développement engagés dans des embryogenèses différentes. L'ensemble des concepts où les microARN matures et les ARNm possèdent cet attribut a été extrait : 93 concepts se répartissent dans une seule composante connexe. Encore une fois, api-mir-3019-5p est impliqué seul dans un très grand nombre de concepts. Afin de simplifier la visualisation et l'analyse, l'ensemble des concepts où api-mir-3019-5p apparaît seul ont été supprimés. On obtient trois composantes connexes avec un, quatre et 59 concepts pour un total de 64 concepts.

La première et deuxième composantes incluent respectivement les microARN matures api-mir-281-5p et api-mir-3026-5p, api-mir-3019-5p. Un seul ARNm dans la seconde composante, ACYPI009164, possède une annotation : « arylsulfatase b-like ». Le peu d'information disponible pour les ARNm de ces composantes rend l'interprétation de ces concepts difficile. La troisième composante, présentée Figure 5.15 et 5.16, fait intervenir api-mir-1-3p, api-mir-87-3p, api-mir-1000-5p, api-mir-novel146-5p, api-mir-263a-5p et api-mir-3019-5p. Certains concepts contiennent les fonctions « cycle cellulaire », « système neuroendocrine », « ovogenèse », « développement musculaire », « épigénétique » et « régulation transcriptionnelle », c'est-à-dire l'ensemble des fonctions sauf « régulation post-transcriptionnelle ». Le concept 20 (Figure 5.16) inclut un ARNm sans annotation ainsi que api-mir-3019-5p, api-mir-1000-5p et api-mir-novel146-5p. La distance entre les sites de fixation de api-mir-1000-5p et api-mir-novel146-5p est de 50 nucléotides ; une coopération entre ces deux sites est donc possible. Le concept 29 (Figure 5.16) inclut l'un des neuf ARNm ciblés par quatre microARN matures, pour rappel le degré le plus élevé observé pour le nombre d'interactions d'un ARNm. Cet ARNm est annoté comme « protein held out wings-like », protéine impliquée dans le développement chez *Drosophila* [170]. Les quatre microARN matures api-mir-3019-5p,

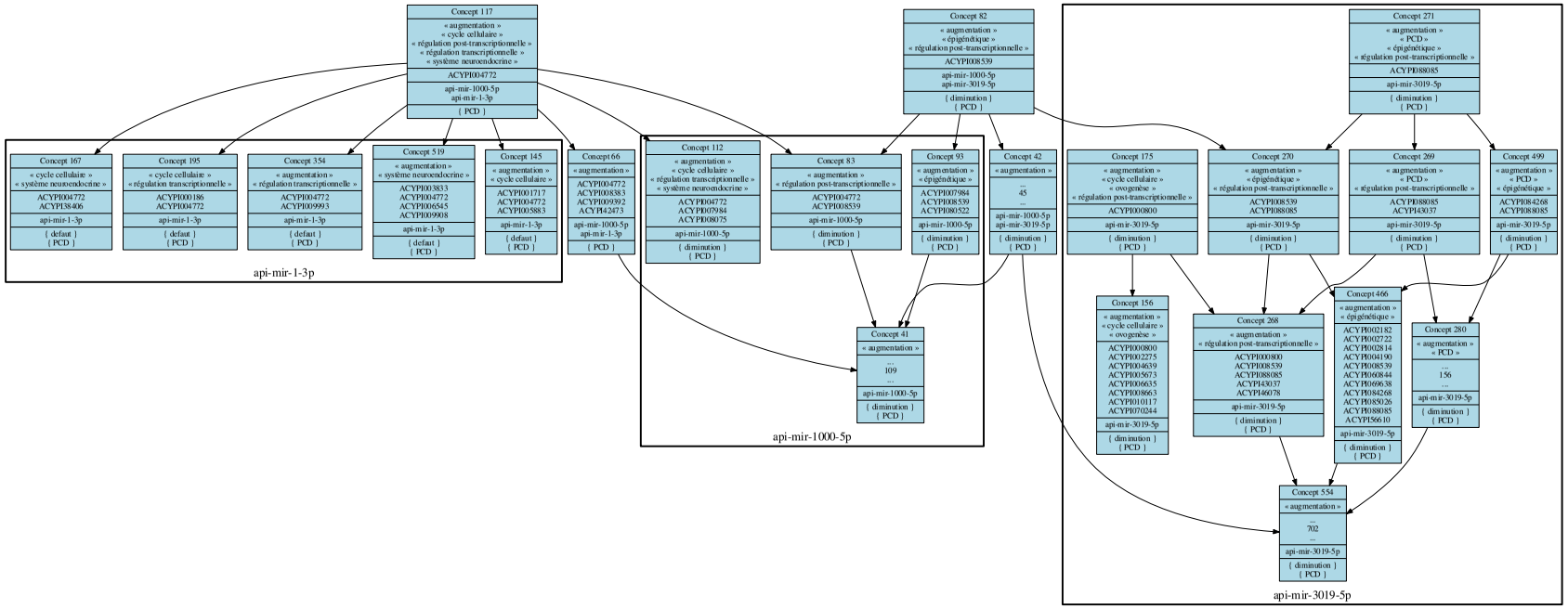


FIG. 5.12 – Deuxième composante connexe impliquant la fonction régulation post-transcriptionnelle. Les concepts encadrés font intervenir un même microARN mature.

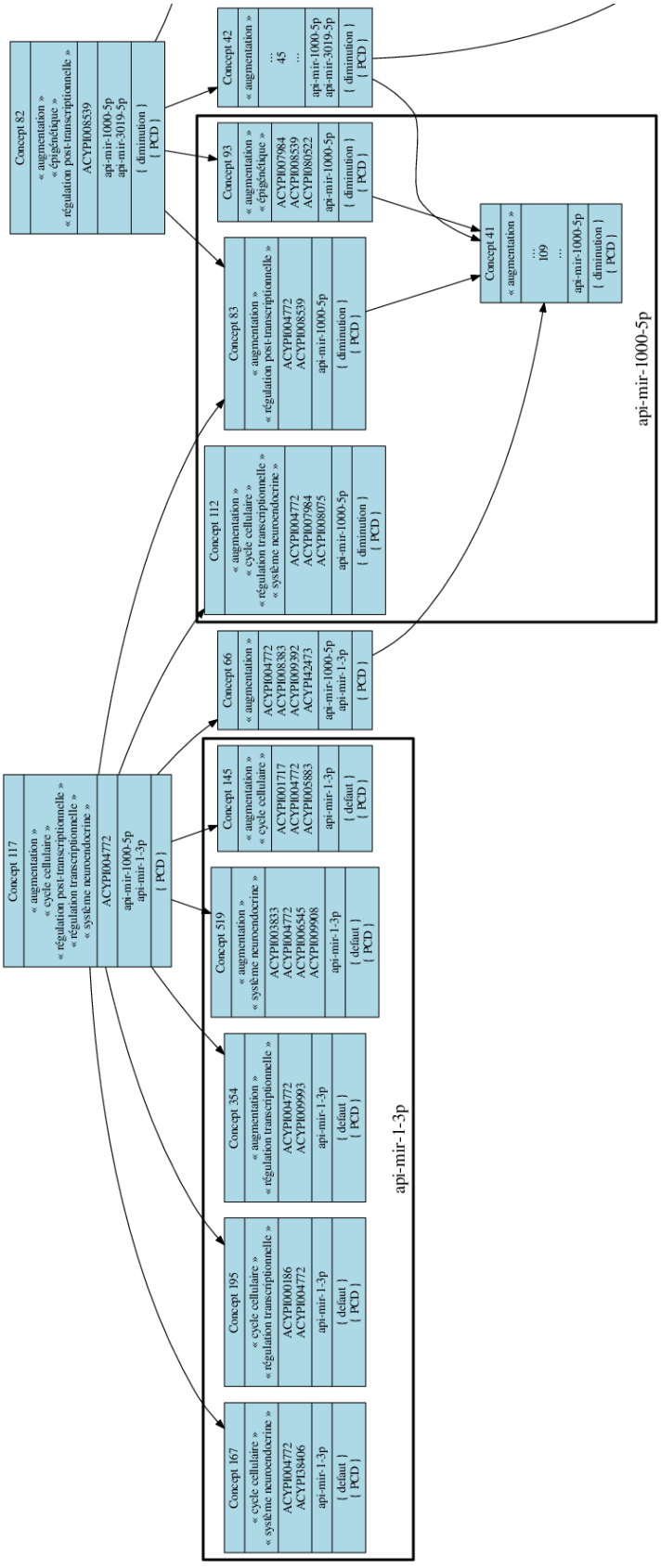


FIG. 5.13 – Zoom sur la partie gauche de la deuxième composante connexe impliquant la fonction régulation post-transcriptionnelle. Les concepts encadrés font intervenir un même microARN mature.

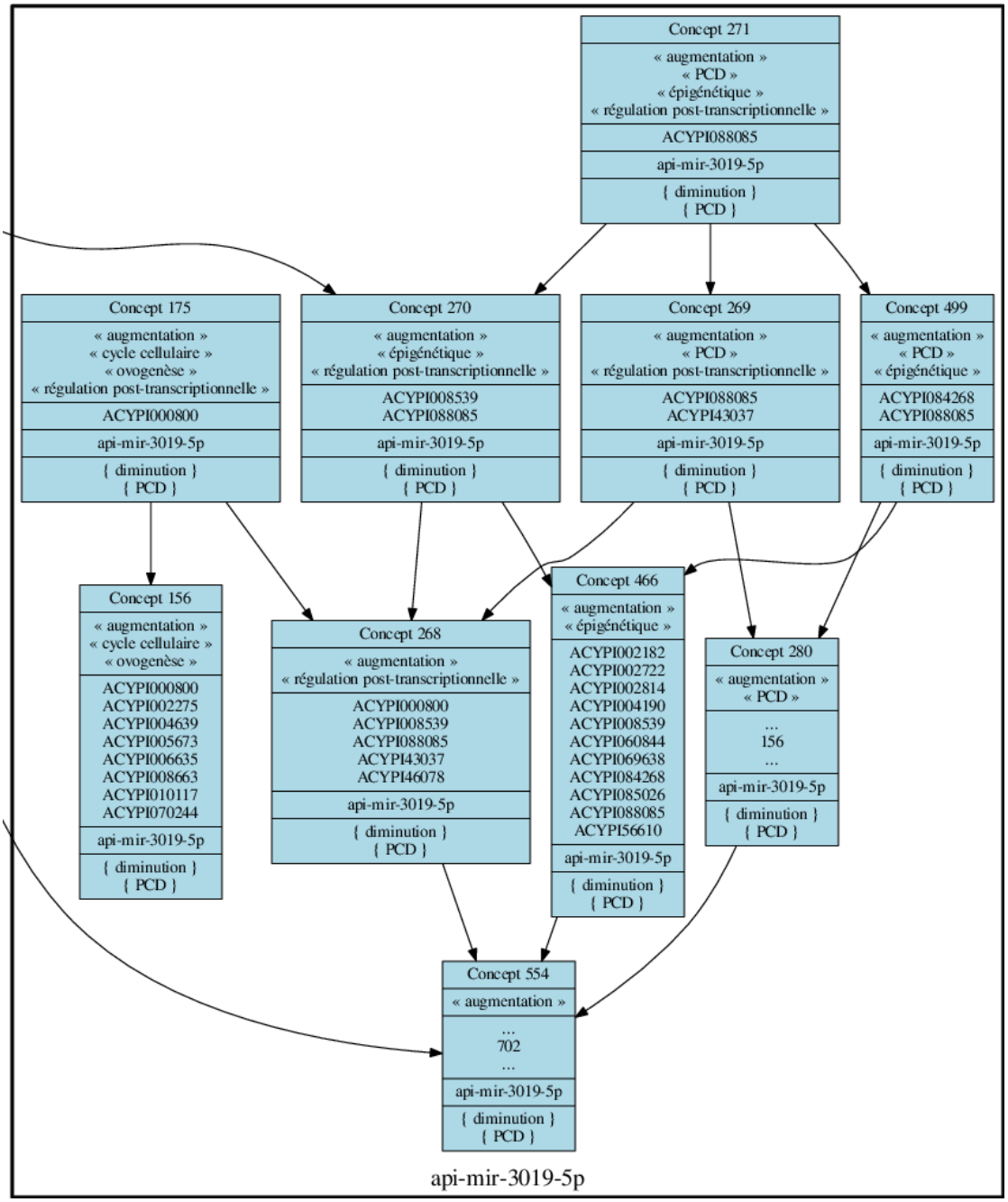


FIG. 5.14 – Zoom sur la partie droite de la deuxième composante connexe impliquant la fonction régulation post-transcriptionnelle. Les concepts encadrés font intervenir un même microARN mature.

api-mir-1000-5p, api-mir-263a-5p et api-mir-novel146-5p possèdent chacun un site de fixation et la distance entre les sites de fixation de api-mir-novel146-5p et api-mir-1000-5p de 68 nucléotides est proche de la limite (60 nucléotides) pour la coopération des sites.

En conclusion, l'analyse du réseau comprenant des éléments régulés entre les deux types d'embryogenèse a été rendue possible grâce à l'enrichissement par des attributs, et par une sélection de sous-réseaux pour faciliter leur interprétation. Des fonctions semblent caractériser ce réseau (ovogenèse précoce, développement...) et des éléments clefs du réseaux ont été identifiés. L'interprétation fine des réseaux et sous-réseaux réclament une annotation et analyse experte afin d'utiliser ces réseaux comme de véritables outils d'aide à la décision. Ceci est présenté en discussion générale.

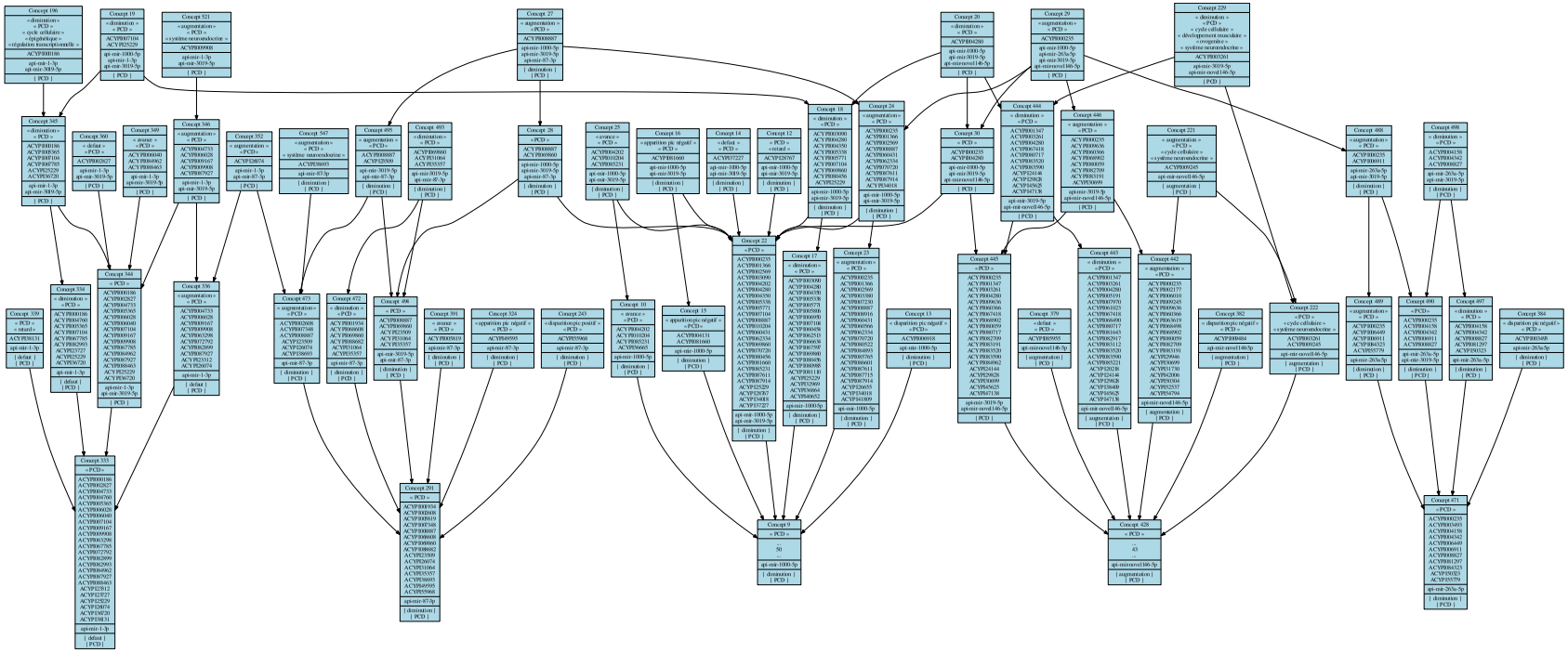


FIG. 5.15 – Troisième composante connexe où l'ensemble des concepts incluent des microARN matures et des ARNm qui possèdent « PCD ».

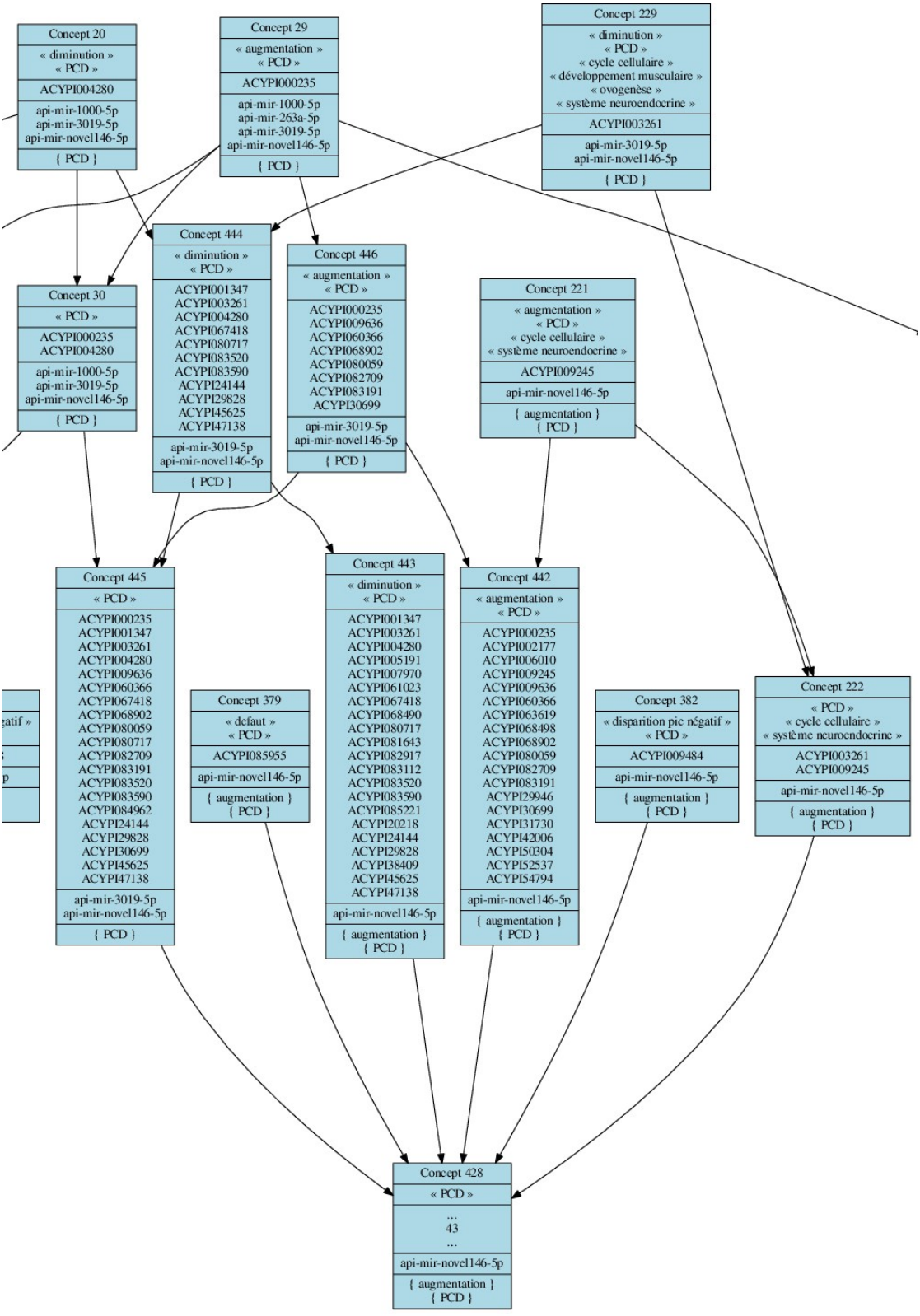


FIG. 5.16 – Zoom sur la partie de la troisième composante connexe où l'ensemble des concepts incluent des microARN matures et des ARNm qui possèdent « PCD » qui contient les concept 20 et 29.

5.3 Résumé et conclusion : étude de modules d'interaction de microARN/ARNm

Lors des chapitres précédents, un réseau d'interactions entre les microARN matures et les ARNm chez *Acyrthosiphon pisum* a été obtenu puis réduit à la question biologique d'intérêt : la discrimination de deux embryogenèses, sexuée et asexuée, par des régulations différentes des ARNm par les microARN matures. Le réseau étant trop important pour pouvoir être étudié manuellement, une méthode de fouille de données à l'aide de l'analyse de concepts formels a été développée. Cette méthode a rendu possible l'ajout au contexte formel représentant les interactions microARN/ARNm, des attributs biologiques concernant ces deux ensembles d'objets. L'ajout d'attributs biologiques permet d'identifier des modules où l'ensemble des microARN matures ou des ARNm partagent tous un ou plusieurs attributs, ce qui aide à l'interprétation de ces modules.

L'énumération des concepts possédant au moins une interaction issus du nouveau contexte a permis l'extraction de 555 concepts dont 65 modules sont caractérisés par des attributs sur les microARN matures et 473 sont caractérisés par des attributs sur les ARNm. Différents critères ont été définis pour fouiller de différentes façons à l'intérieur du treillis formé par ces concepts :

- Les modules où plusieurs microARN matures et ARNm interviennent ;
- Les modules où les interactions correspondent à des règles cohérentes ;
- Les modules avec des fonctions biologiques définies manuellement ;
- Les modules où les microARN matures et/ou les ARNm sont régulés lors du passage de l'embryon d'un état flexible vers un état inflexible (« PCD »).

Pour l'ensemble de ces recherches, différents modules ont pu être identifiés. Certains cas particuliers semblent plus intéressants que d'autres, comme par exemple l'un des sous-graphes sur l'attribut « ovogenèse ». À partir de ce sous-graphe, nous pouvons observer que l'ensemble des ARNm ciblés par le microARN mature api-mir-1000-5p qui possèdent l'attribut « ovogénèse » présentent aussi soit les attributs « cycle cellulaire » et « régulation transcriptionnelle » soit l'attribut « épigénétique ». Cette observation pourrait soutenir le fait que api-mir-1000-5p régulent spécifiquement des ARNm qui sont impliqués dans la régulation transcriptionnelle et le cycle cellulaire lors de l'ovogenèse.

Chapitre 6

Discussion et perspectives

6.1 Discussion générale

Les réseaux de gènes sont utilisés pour comprendre et interpréter les relations qui peuvent exister entre les différents acteurs du fonctionnement du vivant. Ils permettent de visualiser, d'interpréter, de proposer de nouvelles hypothèses et d'aider à la compréhension de processus complexes car impliquant de nombreux éléments et/ou de nombreuses interactions. Modéliser les phénomènes d'interactions en biologie dans le cadre de la théorie des graphes permet d'utiliser le vaste acquis des mathématiques et de l'informatique dans ce domaine. On peut aussi étudier dans un environnement abstrait unifié la caractérisation de phénomènes biologiques tels que la régulation transcriptionnelle ou post-transcriptionnelle, les interactions protéines-protéines ou encore les réactions métaboliques. Nous l'avons utilisé pour notre part pour progresser dans l'étude d'un mécanisme biologique encore peu compris, le polyphénisme de reproduction chez le puceron.

Parmi les processus d'adaptation d'un organisme à son environnement, le polyphénisme, qui est un cas particulier de la plasticité phénotypique, prend une place particulière. Il se définit comme étant la capacité d'un organisme à répondre à des facteurs extérieurs variés par des phénotypes souvent très différenciés les uns des autres afin de faciliter une adaptation au changement de ces facteurs. Plusieurs espèces présentent des caractères dits « plastiques » ce qui implique pour ces organismes des développements et des embryogenèses différentes aboutissant à des phénotypes alternatifs. Ces modes de développement associés à chaque phénotype impliquent que des régulations géniques s'activent lors du développement des individus.

C'est dans ce contexte que cette thèse prend sa place : l'étude d'un réseau biologique appliqué à l'identification de régulations post-transcriptionnelles des ARNm par les microARN lors du polyphénisme du mode de reproduction chez les embryons d'*Acyrthosiphon pisum*, le puceron du pois. C'est dans ce but que différents objectifs ont été proposés :

- Création d'un réseau : mise en relation des microARN matures et des ARNm chez *A. pisum* ;
- Réduction du réseau à la question biologique : caractérisation des microARN matures et des ARNm du réseau ayant des profils d'expression différents selon le type d'embryogenèse ;
- Caractérisation du réseau : identification d'interactions ou de modules d'interactions impliqués dans le caractère plastique du mode de reproduction.

Du point de vue méthodologique, le cœur de notre contribution repose sur l'utilisation de l'analyse de concepts formels (ACF) pour dégager des structures cohérentes dans les données. Ces développements concernent plus particulièrement le cas des graphes bipartis, qui reflètent le caractère bipolaire des interactions.

Dans cette discussion générale, nous reprenons ces différents objectifs.

6.1.1 Création du réseau

Afin d'identifier des régulations post-transcriptionnelles impliquées dans le processus d'intérêt, il faut d'abord décrire ces interactions pangénomiques et donc certains des acteurs de ces interactions : les ARNm et les microARN.

Catalogues des acteurs du réseau : les ARNm et les microARN

L'annotation du génome du puceron du pois en 2010 par IACG [15] et sa mise à jour (disponible sur AphidBase [78]) ont permis l'identification de 36.973 gènes qui transcrivent 36.990 ARNm. C'est ce catalogue qui a été utilisé dans cette thèse. Comparé à d'autres espèces d'insectes, le nombre de transcrits chez *A. pisum* est important : il s'explique notamment par un grand nombre de duplications et d'expansions de familles de gènes [171]. Notons que les gènes codant les protéines de la machinerie de biosynthèse des microARN sont eux-mêmes dupliqués [79].

La même année, le premier catalogue des microARN, comprenant 179 précurseurs et 149 matures, du puceron du pois a été publié par Legeai *et al.* [81]. Ce catalogue a été mis à jour lors de cette thèse à l'aide de données de séquençage haut-débit et a permis d'identifier 329 séquences uniques de pré-microARN qui se répartissent sur 401 gènes de microARN codant pour 573 séquences uniques de microARN matures. Le nombre de gènes de microARN et de microARN matures place *A. pisum* comme une des espèces ayant le plus de microARN parmi son phylum, les Hexapodes. Une des raisons possibles à ce nombre important de microARN est que, comme pour les gènes codant des protéines, de nombreuses duplications ont eu lieu. En effet, l'analyse des gènes de microARN a permis de mettre en évidence des duplications : 30 % (116) des 401 gènes sont dupliqués et représentent 44 séquences de pré-microARN. Ce nombre élevé de duplications est l'une des hypothèses de la présence de clusters de microARN de « grande taille » en comparaison à *Drosophila melanogaster* ainsi que le nombre élevé de gènes dans des familles de microARN chez *A. pisum* comparé à *D. melanogaster*. Les séquences utilisées pour prédire les microARN ont été obtenues sur des embryons en cours de développement, ce qui indique qu'il est possible que de nouveaux microARN soient détectés lors de futurs séquençages d'autres tissus et conditions. De plus, les critères qui ont été utilisés ici pour définir la similarité entre deux microARN matures sont plus stricts que ceux utilisés par miRBase. Cette différence dans la définition de la similarité est à prendre en compte lors de l'observation des résultats concernant les familles de microARN et la comparaison entre *A. pisum* et *D. melanogaster*. Cependant, des erreurs dans la prédiction des microARN ne sont pas à exclure et des confirmations ultérieures, par exemple par RT-PCR, seraient envisageables.

Le réseau

L'annotation des ARNm et des microARN permet la prédiction pangénomique des sites de fixation des microARN matures sur les 3'UTR des ARNm. Les catalogues des 802 microARN matures et des 36.990 ARNm ont permis de prédire, à l'aide de TargetScan v5 [51], le premier réseau d'interactions microARN/ARNm chez *A. pisum* avec : 802 microARN matures, 31.964 ARNm et 1.162.561 sites de fixation ce qui représente 961.915 couples microARN/ARNm. Une comparaison a été menée avec le réseau obtenu chez *D. melanogaster*. Il apparaît que les réseaux diffèrent sur les nombres moyens d'interactions par microARN matures et par ARNm. Nous émettons l'hypothèse que les nombreuses duplications chez *A. pisum* seraient à l'origine de cette différence. Pour cela, il faudrait observer si des gènes issus d'une duplication chez *A. pisum* sont ciblés par les mêmes microARN matures.

En conclusion, le nombre d'interactions microARN/ARNm chez le puceron du pois

est très important. De ces prédictions bio-informatiques, il faut extraire des connaissances pertinentes pour la question biologique posée.

6.1.2 Réduction du réseau à la question biologique

Afin de ne garder que le réseau permettant d'observer les différences entre les embryogenèses sexuées et asexués, il faut le réduire aux interactions d'intérêt. Ces interactions d'intérêt sont définies comme étant celles ayant des formes cinétiques d'expression différentes entre les deux embryogenèses.

Discrétisation des cinétiques

Nous avons discrétisé les cinétiques d'expression pour chacune des embryogenèses prises séparément à la fois pour les microARN matures et pour les ARNm. Cette discrétisation a permis d'associer pour chaque élément un profil de valeurs discrètes pour chaque embryogenèse ; si pour un microARN mature ou un ARNm les deux profils sont différents, alors cela signifie que sa cinétique d'expression est différente entre les deux embryogenèses. L'avantage de cette modélisation des cinétiques est qu'elle permet de définir de façon claire le type d'évolution (et non des points temporels pris indépendamment), de les comparer sur une même base et de les classer en fonction de ces comparaisons. Néanmoins, une perte d'information accompagne la discrétisation. Par exemple, la notion de niveau d'expression faible/forte est perdue. Il se peut aussi que les niveaux d'expression des deux cinétiques se croisent sans que la méthode ne puisse le détecter. Ces paramètres pourront cependant être ajoutés *a posteriori* comme attributs lors de la caractérisation du réseau (voir plus loin).

Nous avons défini un ensemble de règles fournissant une typologie des différences observées entre profils. L'avantage de ces règles est qu'elles permettent de formaliser et de regrouper des différences semblables. Elles permettent de classer les cinétiques en fonction de leur type d'évolution, monotone, transitoire ou décalage temporel. Par contre, la généralisation de ces différences de profils induit une perte d'information sur les temps où ces différences ont lieu. Ces informations peuvent également être ajoutées *a posteriori* comme des attributs, ce que nous avons fait pour la transition de T_0 vers T_1.

Annotation des éléments régulés

Sur la totalité des ARNm, le pourcentage d'ARNm orphelins s'élève à 73 % et sur les ARNm avec des différences de cinétiques d'expression ce nombre passe à 67 %. Deux hypothèses peuvent être faites sur le grand nombre de fonctions inconnues des gènes du puceron du pois : une annotation imparfaite, et des fonctions spécifiques aux pucerons. Cette complexité s'explique aisément si l'on tient compte du fait que les différentes espèces d'insectes pour lesquelles des génomes sont disponibles ont divergé il y a près de 300 millions d'années. Une observation similaire a été constatée pour le génome de la daphnie, un arthropode qui alterne également entre deux modes de reproduction. Ce génome possède un nombre important de gènes orphelins : la plupart de ces gènes orphelins ont des expressions spécifiques dans des conditions de pollution environnementale [172]. Il est donc possible que chez le puceron, les gènes orphelins

correspondent à des fonctions régulées par les changements des environnements locaux, comme ici la photopériode (discuté par Simon *et al.* [173]).

Pour les microARN matures, neuf des 15 microARN ont déjà été identifiés dans d'autres espèces dont sept qui possèdent au moins une annotation connue. Le pourcentage de microARN matures ayant des cinétiques différentes (2,6 %) est faible comparé aux ARNm (13,5 %). Une des raisons possibles de ce faible pourcentage est la moins bonne adéquation du test statistique utilisé qui s'appuie sur les méthodes d'analyse différentielle de la suite de logicielles `edgeR` [124, 125, 126, 127]. Ces méthodes ont été développées spécifiquement pour l'analyse différentielle de l'expression des ARNm. Or ici, notamment dans un souci de cohérence, le même protocole est utilisé pour les ARNm et les microARN matures. On peut se demander si la transposition de ces méthodes aux microARN, séquences de petites tailles, et avec des occurrences très différentes de celles des ARNm est efficace. Notons également que nous avons remarqué qu'en réalisant un regroupement hiérarchique des données d'expression des 21 échantillons de microARN (non montré), les regroupements ne suit pas systématiquement les traitements. Il est donc probable qu'une variabilité forte dans nos données explique en partie le faible nombre de microARN observés. La cause de ces variations n'est pas identifiée : soit une variabilité dans la construction des banques de petits ARN [174], soit une variation faible du niveau de régulation (de l'ordre de 1.5) des microARN observés dans la plupart des modèles biologiques.

Réseau réduit

Une fois les éléments différentiellement régulés identifiés, il est alors possible de réduire le réseau pour en extraire le cœur, c'est-à-dire le réseau formé par les interactions impliquant uniquement ces éléments régulés. Le réseau des éléments régulés implique 15 microARN matures, 1.810 ARNm et 2.795 sites de fixation pour un nombre de couples microARN/ARNm de 2.250.

Sur ce réseau, un fait marquant est le nombre très élevé de cibles de api-mir-3019-5p : 1.300 ARNm sont ciblés par ce microARN matures dont 976 qui ne le sont que par lui. La fonction d'api-mir-3019-5p est inconnue puisque ce microARN n'est pas décrit chez d'autres espèces. On ne peut pas exclure une erreur de prédiction. Néanmoins, nous pouvons aussi poser l'hypothèse que ce microARN mature rentre en compétition avec d'autres microARN matures : son rôle serait d'empêcher, par compétition sur les sites de fixation des 3'UTR, la fixation d'autres microARN matures sur leurs ARNm cibles. Cette compétition aurait lieu si deux sites de fixations sont très proches : la fixation d'un microARN sur l'un des sites empêcherait la fixation de l'autre par encombrement stérique du complexe miRISC. Cette hypothèse pourrait être testée en observant la distance entre les sites de fixation d'api-mir-3019-5p et les autres microARN matures ciblant les mêmes ARNm. Enfin, cette hypothèse reprend en partie la rôle d'éponge que pourraient avoir les microARN dans certaines conditions.

6.1.3 Caractérisation du réseau

Une fois le réseau des éléments régulés obtenu, une analyse est nécessaire pour extraire de la connaissance de ce réseau. Nous avons choisi pour cela d'utiliser l'analyse de concepts formels (ACF). Notre analyse combine l'ACF et l'ajout d'attributs biologiques

pour détecter des modules d'interactions où l'ensemble des microARN matures ou des ARNm partagent des attributs biologiques identiques. Nous en avons extrait un certain nombre d'interprétations visent à extraire de la connaissance de ce réseau pour proposer des ARNm et/ou des microARN ayant des positions ou attributs spéciaux dans le réseau, et pouvant représenter de futurs candidats à caractériser fonctionnellement. Nous reprenons ici les principales observations.

Les modules 36/37 (Figure 5.4) contiennent les deux microARN **api-mir-1000-5p** et **api-mir-263a-5p** tous deux annotés comme étant exprimés dans la tête : soit chez *Apis mellifera*, l'abeille, pour ame-mir-1000 [137] soit chez *Bombyx mori*, bombyx du mûrier, pour bmo-miR-263b. Dans ces mêmes modules, ces deux microARN matures chez *A. pisum* ciblent ACYPI005514, ARNm annoté comme étant impliqué dans le développement du système nerveux périphérique. Or, deux autres ARNm sont présents dans ce module dont **ACYPI002763** qui n'a pas d'annotation. Ce regroupement de deux microARN et d'un ARNm annotés autour de fonctions neuronales permet de poser l'hypothèse que **ACYPI002763** pourrait être aussi en lien avec le développement neuronal ou le système endocrine.

De même, **api-mir-14-3p** et **api-mir-263a-5p** sont au sein du même module 456 (Figure 5.5) ciblant deux ARNm, **ACYPI004009** et **ACYPI008827**, ce dernier étant annoté comme « forkhead box protein o-like » potentiellement impliqué dans l'apoptose [163]. Ces deux microARN matures ont aussi été montrés comme impliqués dans l'apoptose dans des lignées cellulaires de lépidoptères, pour mir-14 [141] et chez *D. melanogaster* pour dme-mir-263a/b [134]. Nous pouvons donc poser l'hypothèse que ACYPI004009 serait lui aussi impliqué dans l'apoptose.

Le sous-graphe de la Figure 5.9 concerne la fonction « ovogenèse ». Le module 77 est constitué de 8 ARNm qui tous partagent les fonctions « cycle cellulaire » et « régulation transcriptionnelle ». Quatre d'entre eux (composant le module 80) possèdent en plus l'annotation « ovogenèse ». Nous proposons que les 8 ARNm du concept 77 soient en partie impliqués dans des fonctions d'ovogenèse, hypothèse à tester par des caractérisations plus fines de l'expression et la fonction de ces transcrits. Un raisonnement similaire peut être appliqué à d'autres concepts du sous-graphe, 23 et 93 sur des fonctions épigénétiques. Ces exemples montrent l'exploitation possible d'un réseau pour proposer de nouvelles hypothèses de fonctions (« coupable par association » »), hypothèses qu'il faut bien sûr tester par des expériences.

Les ARNm **ACYPI008075**, **ACYPI007984** et **ACYPI080522** présents, entre autres, respectivement dans les concepts 108, 116 et 8 peuvent constituer des candidats intéressants pour la caractérisation des deux embryogenèses. En effet, ces trois ARNm se retrouvent ensemble au sein du même concept 81 et partagent les attributs « ovogénèse », « cycle cellulaire », « régulation transcriptionnelle » et « augmentation ». Ils sont respectivement annotés comme « zinc finger protein 1 », « protein mothers against decapentaplegic » (MAD) et « histone-lysine n-methyltransferase eggless-like » (HLME). Ces gènes sont tous trois potentiellement impliqués dans le développement de l'embryon. ACYPI008075 est notamment annoté comme impliqué dans le développement des gonades, ACYPI007984/MAD est une protéine qui réprime l'expression du gène decapentaplegic (gène du développement) chez *D. melanogaster* [175] et ACYPI080522/HLME qui triméthyle Lys-9 sur l'histone H3 dans les ovaires durant l'ovogénèse chez *D. melanogaster* [176].

Au vu des différentes analyses faites, **api-mir-1000-5p** semble être un microARN mature intéressant pour de futures expérimentations biologiques de par son implication dans de nombreux sous-graphes et modules considérés ici comme intéressants, même si il a un nombre important de cibles (197). On le retrouve dans plusieurs concepts et sous-graphes dont celui lié à « ovogenèse » et celui associé à la transition d'un embryon d'un état flexible vers état inflexible (« PCD »). C'est un microARN à fonction inconnue. Nous suggérons en nous basant sur ces observations, que ce microARN pourrait représenter un candidat important à tester fonctionnellement.

Une des difficultés rencontrées pour la caractérisation du réseau est que 8 microARN et 67 % des ARNm n'ont pas d'annotations. L'extraction de processus biologiques impliqués dans le caractère étudié dépend principalement des annotations pré-existantes, que ce soit pour les fonctions des ARNm utilisées dans la création des modules ou pour les annotations des microARN utilisées dans l'interprétation des résultats.

Pour aller plus loin, il faut ajouter de l'information interprétable en termes de régulation. Pour se faire, on pourrait étendre l'analyse en introduisant des attributs biologiques qui n'ont pas été utilisés dans notre étude :

- Le *niveau d'expression* des microARN matures et des ARNm, car la discrétisation « gomme » cet aspect quantitatif. Cet attribut permettrait de révéler dans les réseaux les éléments présentant les plus forts niveaux de régulation ;
- Le *nombre de sites de fixation* pour un couple microARN/ARNm. Plus le nombre de sites de fixation est élevé, plus la répression est potentiellement forte et mérite de s'y intéresser ;
- La *co-localisation génomique* des gènes de microARN et des gènes d'ARNm. Pour les microARN intragéniques, leur expression dépend de celle des ARNm hôtes, et on s'attend à une co-régulation entre microARN et ARNm qui devrait être visible sur le réseau ;
- La *coopération ou la compétition* entre deux sites de fixation. L'effet de la répression est plus important si deux sites sont à la bonne distance.

Du point de vue sémantique, les trois derniers attributs caractérisent non pas les microARN ou les ARNm mais se rapportent aux *interactions* microARN/ARNm. Afin de les intégrer, il faudrait utiliser une méthode permettant de prendre en compte des attributs sur les relations. Une des méthodes qui a été identifiée pendant la thèse est une extension de l'ACF : l'analyse de concepts relationnels (ACR) [177].

L'ACR permet d'analyser des relations entre objets eux mêmes caractérisés par des attributs. Elle utilise plusieurs contextes formels décrivant plusieurs ensembles d'objets par des attributs, par exemple ici les microARN matures, les ARNm et les interactions microARN/ARNm, et des contextes relationnels qui décrivent des relations entre les objets, par exemple ici une interactions microARN/ARNm fait intervenir un ARNm particulier. Nous avons essayé différentes représentations de nos données en ACR mais il s'est montré difficile de pouvoir garder la notion de modules d'interaction développé par notre analyse en ACF en utilisant l'ACR. Néanmoins, arriver à garder cette notion de module permettra de pousser plus loin l'analyse du réseau.

La stratégie employée dans ce travail est basée sur une production exhaustive des modules, sans sélection et élimination a priori de paramètres ou d'éléments (on fonctionne plutôt par ajout d'attributs). Elle nécessite donc la présence d'un expert pour identifier parmi les différents modules ceux qui lui semblent particulièrement liés au

processus étudié. Par contre, l'avantage d'une telle stratégie est l'exhaustivité de l'information rendue disponible par la méthode.

6.1.4 L'analyse de concepts formels dans le contexte d'un graphe biparti

En plus de l'utilisation de l'ACF pour analyser et ajouter de l'information biologique au réseau, l'ACF a été utilisée pour réparer des contextes formels bruités ainsi que pour visualiser des graphes bipartis par des bicliques. La réparation se base sur l'hypothèse d'un regroupement initial en modules du réseau d'interactions microARN/ARNm. Les différents tests réalisés sur les contextes formels bruités mériteraient d'être effectués à plus grande échelle sur des contextes de tailles plus élevées. Les résultats obtenus sur les simulations des graphes d'interactions microARN/ARNm sont moins bons que ceux qui ont été obtenus sur les contextes bruités. Ces mauvais résultats peuvent venir de différents facteurs comme notamment une extraction des paramètres du réseau biologique trop grossière qui n'a pas permis de prendre en compte la vraie topologie sous-jacente du réseau.

La deuxième méthode développée est une méthode de visualisation de graphes bipartis. Cette méthode se place dans le contexte de la couverture optimale d'un graphe biparti par ses bicliques afin de minimiser le nombre d'arêtes représentées. Dans le cadre de l'ACF nous avons proposé une méthode, inspirée de celle développée par Royer *et al.* [154] (Power Graph), permettant de visualiser un graphe biparti en s'aidant de ses concepts et sous-concepts. La méthode ne permet pas encore d'avoir une réduction des arêtes aussi forte que la méthode dont elle s'inspire. Néanmoins, le réel intérêt de cette méthode réside dans la spécification de la visualisation souhaitée. En effet, elle permet de regrouper les nœuds du graphe selon différents critères, contrairement à la méthode Power Graph qui n'optimise que le recouvrement des arêtes. Grâce à notre méthode, la définition de différentes optimisations permet de générer différentes visualisations qui répondent à des questions biologiques différentes.

En conclusion, cette thèse s'applique à analyser l'un des modes de régulation potentiellement impliqué dans le polyphénisme de reproduction du puceron : l'inhibition de la traduction et la dégradation des ARNm induite par les microARN. Cette première analyse a permis de développer les bases d'une modélisation des résultats pouvant être obtenus à l'aide du protocole biologique expérimental sur lequel se base ce travail. Une méthode de fouille de données d'interactions à l'aide d'attributs biologiques a été introduite permettant d'aider à la caractérisation d'interactions clés dans le processus étudié.

6.2 Conclusion et Perspectives

6.2.1 Conclusion

Dans cette thèse, nous avons créé le premier réseau pangénomique représentant le potentiel de l'inhibition de la traduction des ARNm par les microARN chez *Acyrthosiphon pisum*. Ce réseau a été réduit à 2.250 interactions entre 15 microARN matures et 1.810 ARNm ayant des cinétiques différentes entre deux embryogenèses, une sexuée et une asexuée. L'application de l'analyse de concepts formels à ce réseau d'interactions a permis l'identification de différentes fonctions potentiellement importantes dans la discrimination de ces embryogenèses comme l'ovogenèse, la régulation transcriptionnelle ou encore le système neuroendocrine. L'obtention de modules d'interactions respectant certaines contraintes a aussi permis d'identifier des cibles plus précises, comme le microARN mature api-mir-1000-5p qui semble réguler de nombreux modules d'intérêt. Ce travail permet donc de s'approprier les approches de biologie des systèmes pour inférer des fonctions biologiques potentiellement impliquées dans la plasticité phénotypique du mode de reproduction chez un insecte. Il permet de dégager de nouvelles perspectives qui sont développées ci-dessous.

6.2.2 Perspectives

Développer les méthodes liées à l'utilisation de l'analyse de concepts formels

Nous avons effectué trois travaux sur l'utilisation de l'ACF dans le cadre d'analyses sur des réseaux bipartis : la réparation ; la visualisation ; l'analyse. Ces trois travaux mériteraient des développements plus poussés. Pour la réparation, le travail effectué ici a permis une première formalisation de l'influence du bruit au sein d'un treillis de concepts ainsi que l'utilisation des concepts. Néanmoins, une formalisation du résultat de la suppression et de l'ajout d'interactions sur le treillis de concepts devrait permettre d'améliorer le processus de réparation. Ce processus appliqué aux prédictions d'interactions ne pend pas en compte la valeur de score de l'interaction prédite. Une modification de ce processus pour prendre en compte ce score, comme la minimisation des scores des interactions ajoutées et la maximisation de celles supprimées, pourrait améliorer les résultats.

La visualisation ne couvre pas encore l'ensemble des arêtes. Là encore, une étude plus poussée des relations existants entre les ensembles des concepts et sous-concepts permettrait de réduire encore plus le réseau. Nous avons pu voir que les visualisations générées pouvaient répondre à différents critères d'optimisation. Cet aspect pourrait être utilisé dans le cadre de l'analyse d'un réseau par ajout d'information : la méthode de visualisation optimiserait le regroupement des interactions en fonction des concepts obtenus par ajout d'information.

Étendre l'ensemble des attributs utilisés

Afin d'effectuer une recherche et une identification plus précises des modules impliqués dans le polyphénisme de reproduction, il faut ajouter de nouveaux attributs caractérisant les interactions elles-mêmes, comme le nombre de sites de fixation ou encore la distance entre deux sites de fixation. Afin d'ajouter ces attributs à l'analyse,

l'analyse de concepts relationnels (ACR) a été identifiée comme étant une méthode adaptée. Après différents essais pour modéliser la problématique d'intérêt à l'aide de l'ACR, il est apparu que la conservation des modules initiaux, c'est-à-dire des concepts formels liant microARN et ARNm, était délicate à prendre en compte. De futurs travaux devraient permettre la modélisation par l'ACR affin d'approfondir l'analyse des interactions.

Rendre accessible les différents résultats

L'ensemble des résultats présents ici ne sont pas encore pleinement disponibles à la communauté biologique. Un travail d'ingénierie est encore nécessaire afin de fournir à l'ensemble des personnes travaillant sur ce caractère chez le puceron un accès simple aux données et aux résultats des différentes étapes effectuées. Pour cela, un outil bioinformatique devra être développé pour pouvoir afficher et naviguer au sein du treillis de concepts. Des requêtes pourront être effectuées sur ce treillis pour extraire automatiquement des sous-graphes de la même façon que ceux obtenus ici. De plus, cet outil devra permettre d'accéder facilement aux différentes données disponibles comme les valeurs d'expressions, les règles ou encore les annotations directement à partir du treillis.

Analyse fonctionnelle

Comme il a été dit dans la discussion générale, la fouille de données à l'aide de l'ACF génère l'ensemble des modules respectant les conditions fixées par l'analyste. Afin d'identifier plus précisément les microARN et ARNm impliqués dans la plasticité phénotypique de reproduction du puceron du pois, l'exploration facilitée pour un expert de l'ensemble des modules générés est nécessaire. Après cette phase, une caractérisation *in vivo* des éléments identifiés doit être menée pour valider ou non le caractère fonctionnel de ces éléments. L'analyse fonctionnelle peut être faite par ARN interférent (ARNi) pour les ARNm. On peut également procéder par inhibition des microARN, soit en utilisant un antagomir, un oligonucléotide anti-sens au microARN mature, soit en utilisant une éponge à microARN, c'est-à-dire une séquence transfectée dans l'organisme possédant des sites de fixation pour un microARN mature d'intérêt. Une des difficultés majeur de ces expérimentations est que l'ensemble de ces méthodes génétiques doivent être mises au point pour pouvoir passer la barrière embryonnaire afin d'avoir un effet sur l'embryon lors de l'injection chez la mère.

L'ARNi fonctionne chez le puceron du pois mais dans certaines conditions (essentiellement pour des gènes exprimés dans le tube digestif ou les glandes salivaires). Des nouvelles techniques de transgénèse sont en cours de test (TALEN, CRISPR) mais restent difficiles à mettre en œuvre. La recherche de mutants est envisageable [6]. Cependant, en absence aujourd'hui de méthodes d'analyse fonctionnelle routinières, des caractérisations plus précises de l'expression des microARN et des ARNm candidats peuvent être réalisées (q-RT-PCR, hybridation *in situ*). De plus, il est possible de comparer ces expressions entre différentes lignées de puceron du pois qui varient selon leur réponse à la photopériode : certaines sont incapables de changer de mode de reproduction en condition de photopériode courte. Ces lignées représentent un matériel génétique de choix pour affiner les connaissances sur le fonctionnements des gènes de candidats du réseau.

Étude de nouvelles sources de régulations

De nombreux types de régulation existent au sein du vivant et la régulation post-transcriptionnelle des ARNm par les microARN n'est qu'une régulation parmi d'autres. Cette étude ne prend pas en compte d'autres types de régulations comme les régulations épigénétiques ou transcriptionnelles pour ne citer qu'elles. Il a été montré que l'ensemble de ces éléments pouvaient interagir entre eux comme les boucles de régulations microARN, facteurs de transcription et ARNm [167], l'effet tampon des ARNm et des longs ARN non codant sur les microARN [43, 44, 45, 46] ou encore la régulation des ARNm par les longs ARN non codant [178, 179, 180]. Au sein de l'équipe Écologie et Génétique des Insectes à l'IGEPP, les lARNnc et les ARNpi ont été identifiés à l'aide des jeux de données sur le séquençage des longs et petits ARN présentés dans cette thèse. En plus de ces données déjà disponibles, des données épigénétiques sur les changements dans l'accessibilité de la chromatine sont en cours d'acquisition sur le même protocole qui est décrit ici. L'extension du réseau à la diversité des différents phénomènes de régulations connus permettra donc de pousser plus loin la compréhension du changement de mode de reproduction chez *Acyrthosiphon pisum* [6].

Utilisation du réseau pangénomique

Le réseau initial d'interaction entre microARN matures et ARNm que nous avons construit est un réseau pangénomique. Il peut donc être appliqué à d'autres traits de vie d'intérêt chez le puceron du pois comme par exemple la plasticité phénotypique de dispersion (pucerons ailés ou non) ou encore l'adaptation à la plante hôte. Ce réseau permettra à l'ensemble de la communauté travaillant sur le puceron du pois (IAGC) de pouvoir analyser et interpréter des données d'expression de microARN et d'ARNm dans le cadre de la régulation post-transcriptionnelle des gènes.

Bibliographie

[1] Kwang-Il Goh, Michael E Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-László Barabási. The human disease network. *Proceedings of the National Academy of Sciences of the United States of America*, 104(21) :8685–90, May 2007.

[2] François Jacob, D Perrin, C Sanchez, and Jacques Monod. L'opéron : groupe de gènes à expresion coordonnée par un opérateur . *Comptes Rendus Academie des Sciences Paris*, 250 :1727–1729 ST – L'opéron : groupe de gènes à expres, 1960.

[3] Truman P Young, Maureen L Stanton, and Caroline E Christian. Effects of natural and simulated herbivory on spine lengths of Acacia drepanolobium in Kenya. *Oikos*, 101 :171–179, 2003.

[4] E Greene. A diet-induced developmental polymorphism in a caterpillar. *Science (New York, N.Y.)*, 243(4891) :643–6, February 1989.

[5] John D Hatle. Physiology underlying phenotypic plasticity and polyphenisms : introduction to the symposium. *Integrative and comparative biology*, 43(5) :605–6, November 2003.

[6] Denis Tagu, John K Colbourne, and Nicolas Nègre. Genomic data integration for ecological and evolutionary traits in non-model organisms. *BMC genomics*, 15(1) :490, January 2014.

[7] R C Lee, R L Feinbaum, and V Ambros. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75(5) :843–54, December 1993.

[8] B Wightman, I Ha, and G Ruvkun. Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. *Cell*, 75(5) :855–62, December 1993.

[9] Maria I Almeida, Rui M Reis, and George A Calin. MicroRNA history : discovery, recent applications, and next frontiers. *Mutation research*, 717(1-2) :1–8, December 2011.

[10] B J Reinhart, F J Slack, M Basson, A E Pasquinelli, J C Bettinger, A E Rougvie, H R Horvitz, and G Ruvkun. The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature*, 403(6772) :901–6, February 2000.

[11] A E Pasquinelli, B J Reinhart, F Slack, M Q Martindale, M I Kuroda, B Maller, D C Hayward, E E Ball, B Degnan, P Müller, J Spring, A Srinivasan, M Fishman, J Finnerty, J Corbo, M Levine, P Leahy, E Davidson, and G Ruvkun. Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature*, 408(6808) :86–9, November 2000.

[12] M Lagos-Quintana, R Rauhut, W Lendeckel, and T Tuschl. Identification of novel genes coding for small expressed RNAs. *Science (New York, N.Y.)*, 294 :853–858, 2001.

[13] N C Lau, L P Lim, E G Weinstein, and D P Bartel. An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science (New York, N.Y.)*, 294 :858–862, 2001.

[14] R C Lee and V Ambros. An extensive class of small RNAs in Caenorhabditis elegans. *Science (New York, N.Y.)*, 294 :862–864, 2001.

[15] The International Aphid Genomics Consortium. Genome sequence of the pea aphid Acyrthosiphon pisum. *PLoS biology*, 8(2) :e1000313, February 2010.

[16] David P Bartel. MicroRNAs : target recognition and regulatory functions. *Cell*, 136(2) :215–33, January 2009.

[17] Keira Lucas and Alexander S Raikhel. Insect microRNAs : biogenesis, expression profiling and biological functions. *Insect biochemistry and molecular biology*, 43(1) :24–38, January 2013.

[18] Qinghua Jiang, Yadong Wang, Yangyang Hao, Liran Juan, Mingxiang Teng, Xinjun Zhang, Meimei Li, Guohua Wang, and Yunlong Liu. miR2Disease : a manually curated database for microRNA deregulation in human disease. *Nucleic acids research*, 37(Database issue) :D98–104, January 2009.

[19] Sam Griffiths-Jones. The microRNA Registry. *Nucleic acids research*, 32(Database issue) :D109–11, January 2004.

[20] Sam Griffiths-Jones, Russell J Grocock, Stijn van Dongen, Alex Bateman, and Anton J Enright. miRBase : microRNA sequences, targets and gene nomenclature. *Nucleic acids research*, 34(Database issue) :D140–4, January 2006.

[21] Sam Griffiths-Jones, Harpreet Kaur Saini, Stijn van Dongen, and Anton J Enright. miRBase : tools for microRNA genomics. *Nucleic acids research*, 36(Database issue) :D154–8, January 2008.

[22] Ana Kozomara and Sam Griffiths-Jones. miRBase : integrating microRNA annotation and deep-sequencing data. *Nucleic acids research*, 39(Database issue) :D152–7, January 2011.

[23] Ana Kozomara and Sam Griffiths-Jones. mirbase : annotating high confidence micrornas using deep sequencing data. *Nucleic Acids Research*, 42(D1) :D68–D73, 2014.

[24] Clarissa P C Gomes, Ji-Hoon Cho, Leroy Hood, Octávio L Franco, Rinaldo W Pereira, and Kai Wang. A Review of Computational Tools in microRNA Discovery. *Frontiers in genetics*, 4 :81, January 2013.

[25] Masafumi Nozawa, Sayaka Miura, and Masatoshi Nei. Origins and evolution of microRNA genes in Drosophila species. *Genome biology and evolution*, 2 :180–9, January 2010.

[26] Matthew A Saunders, Han Liang, and Wen-Hsiung Li. Human polymorphism at microRNAs and microRNA target sites. *Proceedings of the National Academy of Sciences of the United States of America*, 104(9) :3300–5, February 2007.

[27] Yoontae Lee, Minju Kim, Jinju Han, Kyu-Hyun Yeom, Sanghyuk Lee, Sung Hee Baek, and V Narry Kim. MicroRNA genes are transcribed by RNA polymerase II. *The EMBO journal*, 23(20) :4051–60, October 2004.

[28] Y. Lee. MicroRNA maturation : stepwise processing and subcellular localization. *The EMBO Journal*, 21(17) :4663–4670, September 2002.

[29] Rui Yi, Yi Qin, Ian G Macara, and Bryan R Cullen. Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. *Genes & development*, 17(24) :3011–6, December 2003.

[30] Michele Trabucchi, Paola Briata, Mariaflor Garcia-Mayoral, Astrid D Haase, Witold Filipowicz, Andres Ramos, Roberto Gherzi, and Michael G Rosenfeld. The RNA-binding protein KSRP promotes the biogenesis of a subset of microRNAs. *Nature*, 459(7249) :1010–4, June 2009.

[31] Julia Winter, Stephanie Jung, Sarina Keller, Richard I Gregory, and Sven Diederichs. Many roads to maturity : microRNA biogenesis pathways and their regulation. *Nature cell biology*, 11(3) :228–34, March 2009.

[32] Angélique Girard, Ravi Sachidanandam, Gregory J Hannon, and Michelle A Carmell. A germline-specific class of small RNAs binds mammalian Piwi proteins. *Nature*, 442(7099) :199–202, July 2006.

[33] Katsutomo Okamura, Na Liu, and Eric C Lai. Distinct mechanisms for microRNA strand selection by Drosophila Argonautes. *Molecular cell*, 36(3) :431–44, November 2009.

[34] Katsutomo Okamura, Michael D Phillips, David M Tyler, Hong Duan, Yu-ting Chou, and Eric C Lai. The regulatory activity of microRNA* species has substantial influence on microRNA and 3' UTR evolution. *Nature structural & molecular biology*, 15(4) :354–63, April 2008.

[35] Jr-Shiuan Yang, Michael D Phillips, Doron Betel, Ping Mu, Andrea Ventura, Adam C Siepel, Kevin C Chen, and Eric C Lai. Widespread regulatory activity of vertebrate microRNA* species. *RNA (New York, N.Y.)*, 17(2) :312–26, February 2011.

[36] Antony Rodriguez, Sam Griffiths-Jones, Jennifer L Ashurst, and Allan Bradley. Identification of mammalian microRNA host genes and transcription units. *Genome research*, 14(10A) :1902–10, October 2004.

[37] Noam Shomron and Carmit Levy. MicroRNA-biogenesis and Pre-mRNA splicing crosstalk. *Journal of biomedicine & biotechnology*, 2009 :594678, January 2009.

[38] Helen J Curtis, Christopher R Sibley, and Matthew JA Wood. Mirtrons, an emerging class of atypical mirna. *Wiley Interdisciplinary Reviews : RNA*, 3(5) :617–632, 2012.

[39] Marianthi Kiriakidou, Grace S Tan, Styliani Lamprinaki, Mariangels De Planell-Saguer, Peter T Nelson, and Zissimos Mourelatos. An mRNA m7G cap binding-like motif within human Ago2 represses translation. *Cell*, 129(6) :1141–51, June 2007.

[40] Isabelle Behm-Ansmant, Jan Rehwinkel, Tobias Doerks, Alexander Stark, Peer Bork, and Elisa Izaurralde. mRNA degradation by miRNAs and GW182 requires

both CCR4 :NOT deadenylase and DCP1 :DCP2 decapping complexes. *Genes & development*, 20(14) :1885–98, July 2006.

[41] Sergej Djuranovic, Ali Nahvi, and Rachel Green. miRNA-mediated gene silencing by translational repression followed by mRNA deadenylation and decay. *Science (New York, N.Y.)*, 336(6078) :237–40, April 2012.

[42] På l Saetrom, Bret S E Heale, Ola Snø ve, Lars Aagaard, Jessica Alluin, and John J Rossi. Distance constraints between microRNA target sites dictate efficacy and cooperativity. *Nucleic acids research*, 35(7) :2333–42, January 2007.

[43] Reena V. Kartha and Subbaya Subramanian. Competing endogenous RNAs (ceRNAs) : new entrants to the intricacies of gene regulation. *Frontiers in Genetics*, 5 :8, January 2014.

[44] Rituparno Sen, Suman Ghosal, Shaoli Das, Subrata Balti, and Jayprokas Chakrabarti. Competing endogenous RNA : the key to posttranscriptional regulation. *TheScientificWorldJournal*, 2014 :896206, January 2014.

[45] Hervé Seitz. Redefining microRNA targets. *Current biology : CB*, 19(10) :870–3, May 2009.

[46] Leonardo Salmena, Laura Poliseno, Yvonne Tay, Lev Kats, and Pier Paolo Pandolfi. A ceRNA hypothesis : the Rosetta Stone of a hidden RNA language ? *Cell*, 146(3) :353–8, August 2011.

[47] T M Witkos, E Koscianska, and W J Krzyzosiak. Practical Aspects of microRNA Target Prediction. *Current molecular medicine*, 11(2) :93–109, March 2011.

[48] Hao Zheng, Rongguo Fu, Jin-Tao Wang, Qinyou Liu, Haibin Chen, and Shi-Wen Jiang. Advances in the Techniques for the Prediction of microRNA Targets. *International journal of molecular sciences*, 14(4) :8179–87, January 2013.

[49] Sarah M Peterson, Jeffrey A Thompson, Melanie L Ufkin, Pradeep Sathyanarayana, Lucy Liaw, and Clare Bates Congdon. Common features of microRNA target prediction tools. *Frontiers in genetics*, 5 :23, January 2014.

[50] Kyle Kai-How Farh, Andrew Grimson, Calvin Jan, Benjamin P Lewis, Wendy K Johnston, Lee P Lim, Christopher B Burge, and David P Bartel. The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. *Science (New York, N.Y.)*, 310 :1817–1821, 2005.

[51] Andrew Grimson, Kyle Kai-How Farh, Wendy K Johnston, Philip Garrett-Engele, Lee P Lim, and David P Bartel. Microrna targeting specificity in mammals : determinants beyond seed pairing. *Molecular cell*, 27(1) :91–105, 2007.

[52] Michael Kertesz, Nicola Iovino, Ulrich Unnerstall, Ulrike Gaul, and Eran Segal. The role of site accessibility in microRNA target recognition. *Nature genetics*, 39(10) :1278–84, October 2007.

[53] Doron Betel, Anjali Koppal, Phaedra Agius, Chris Sander, and Christina Leslie. Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome biology*, 11(8) :R90, January 2010.

[54] Martin Reczko, Manolis Maragkakis, Panagiotis Alexiou, Ivo Grosse, and Artemis G Hatzigeorgiou. Functional microRNA targets in protein coding sequences. *Bioinformatics (Oxford, England)*, 28(6) :771–6, March 2012.

[55] Maria D Paraskevopoulou, Georgios Georgakilas, Nikos Kostoulas, Ioannis S Vlachos, Thanasis Vergoulis, Martin Reczko, Christos Filippidis, Theodore Dalamagas, and A G Hatzigeorgiou. DIANA-microT web server v5.0 : service integration into miRNA functional analysis workflows. *Nucleic acids research*, 41(Web Server issue) :W169–73, July 2013.

[56] Shiping Liu, Song Gao, Danyu Zhang, Jiyun Yin, Zhonghuai Xiang, and Qingyou Xia. MicroRNAs show diverse and dynamic expression patterns in multiple tissues of Bombyx mori. *BMC genomics*, 11(1) :85, January 2010.

[57] Anton J Enright, Bino John, Ulrike Gaul, Thomas Tuschl, Chris Sander, and Debora S Marks. MicroRNA targets in Drosophila. *Genome biology*, 5(1) :R1, January 2003.

[58] Benjamin P Lewis, Christopher B Burge, and David P Bartel. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, 120(1) :15–20, January 2005.

[59] Marianthi Kiriakidou, Peter T Nelson, Andrei Kouranov, Petko Fitziev, Costas Bouyioukos, Zissimos Mourelatos, and Artemis Hatzigeorgiou. A combined computational-experimental approach predicts human microRNA targets. *Genes & development*, 18(10) :1165–78, May 2004.

[60] Azra Krek, Dominic Grün, Matthew N Poy, Rachel Wolf, Lauren Rosenberg, Eric J Epstein, Philip MacMenamin, Isabelle da Piedade, Kristin C Gunsalus, Markus Stoffel, and Nikolaus Rajewsky. Combinatorial microRNA target predictions. *Nature genetics*, 37 :495–500, 2005.

[61] Sabbi Lall, Dominic Grün, Azra Krek, Kevin Chen, Yi Lu Wang, Colin N. Dewey, Pranidhi Sood, Teresa Colombo, Nicolas Bray, Philip MacMenamin, Huey Ling Kao, Kristin C. Gunsalus, Lior Pachter, Fabio Piano, and Nikolaus Rajewsky. A genome-wide map of conserved MicroRNA targets in C. elegans. *Current Biology*, 16 :460–471, 2006.

[62] Kevin C. Miranda, Tien Huynh, Yvonne Tay, Yen Sin Ang, Wai Leong Tam, Andrew M. Thomson, Bing Lim, and Isidore Rigoutsos. A Pattern-Based Method for the Identification of MicroRNA Binding Sites and Their Corresponding Heteroduplexes. *Cell*, 126 :1203–1217, 2006.

[63] Praveen Sethupathy, Molly Megraw, and Artemis G Hatzigeorgiou. A guide through present computational approaches for the identification of mammalian microRNA targets. *Nature methods*, 3(11) :881–6, November 2006.

[64] Praveen Sethupathy, Benoit Corda, and Artemis G Hatzigeorgiou. TarBase : A comprehensive database of experimentally supported animal microRNA targets. *RNA (New York, N.Y.)*, 12(2) :192–7, February 2006.

[65] Daehyun Baek, Judit Villén, Chanseok Shin, Fernando D Camargo, Steven P Gygi, and David P Bartel. The impact of microRNAs on protein output. *Nature*, 455(7209) :64–71, September 2008.

[66] Panagiotis Alexiou, Manolis Maragkakis, Giorgos L Papadopoulos, Martin Reczko, and Artemis G Hatzigeorgiou. Lost in translation : an assessment and perspective for computational microRNA target identification. *Bioinformatics (Oxford, England)*, 25(23) :3049–55, December 2009.

[67] Qinghua Cui, Zhenbao Yu, Enrico O Purisima, and Edwin Wang. Principles of microRNA regulation of a human cellular signaling network. *Molecular systems biology*, 2(1) :46, January 2006.

[68] Kang Tu, Hui Yu, You-Jia Hua, Yuan-Yuan Li, Lei Liu, Lu Xie, and Yi-Xue Li. Combinatorial network of primary and secondary microRNA-driven regulatory mechanisms. *Nucleic acids research*, 37(18) :5969–80, October 2009.

[69] Bing Liu, Jiuyong Li, and Murray J Cairns. Identifying miRNAs, targets and functions. *Briefings in bioinformatics*, 15(1) :1–19, January 2014.

[70] Kenneth Bryan, Marta Terrile, Isabella M Bray, Raquel Domingo-Fernandéz, Karen M Watters, Jan Koster, Rogier Versteeg, and Raymond L Stallings. Discovery and visualization of miRNA-mRNA functional modules within integrated data using bicluster analysis. *Nucleic acids research*, 42(3) :e17, February 2014.

[71] GAEL LE TRIONNAIRE, VALENTIN WUCHER, and DENIS TAGU. Genome expression control during the photoperiodic response of aphids. *Physiological Entomology*, 38(2) :117–125, June 2013.

[72] CG Steel and AD Lees. The role of neurosecretion in the photoperiodic control of polymorphism in the aphid Megoura viciae. *J. Exp. Biol.*, 67(1) :117–135, April 1977.

[73] G Le Trionnaire, F Francis, S Jaubert-Possamai, J Bonhomme, E De Pauw, J-P Gauthier, E Haubruge, F Legeai, N Prunier-Leterme, J-C Simon, S Tanguy, and D Tagu. Transcriptomic and proteomic analyses of seasonal photoperiodism in the pea aphid. *BMC genomics*, 10(1) :456, January 2009.

[74] J Huybrechts, J Bonhomme, S Minoli, N Prunier-Leterme, A Dombrovsky, M Abdel-Latief, A Robichon, J A Veenstra, and D Tagu. Neuropeptide and neurohormone precursors in the pea aphid, Acyrthosiphon pisum. *Insect molecular biology*, 19 Suppl 2 :87–95, March 2010.

[75] Terry S. Corbitt and Jim Hardie. Juvenile hormone effects on polymorphism in the pea aphid, A cyrthosiphon pisum. *Entomologia Experimentalis et Applicata*, 38(2) :131–135, July 1985.

[76] Toru Miura, Christian Braendle, Alexander Shingleton, Geoffroy Sisk, Srinivas Kambhampati, and David L Stern. A comparison of parthenogenetic and sexual embryogenesis of the pea aphid Acyrthosiphon pisum (Hemiptera : Aphidoidea). *Journal of experimental zoology. Part B, Molecular and developmental evolution*, 295(1) :59–81, February 2003.

[77] Aurore Gallot, Shuji Shigenobu, Tomomi Hashiyama, Stéphanie Jaubert-Possamai, and Denis Tagu. Sexual and asexual oogenesis require the expression of unique and shared sets of genes in the insect Acyrthosiphon pisum. *BMC genomics*, 13(1) :76, January 2012.

[78] F Legeai, S Shigenobu, J-P Gauthier, J Colbourne, C Rispe, O Collin, S Richards, A C C Wilson, T Murphy, and D Tagu. AphidBase : a centralized bioinformatic resource for annotation of the pea aphid genome. *Insect molecular biology*, 19 Suppl 2 :5–12, March 2010.

[79] Stéphanie Jaubert-Possamai, Claude Rispe, Sylvie Tanguy, Karl Gordon, Thomas Walsh, Owain Edwards, and Denis Tagu. Expansion of the miRNA pathway in

the hemipteran insect Acyrthosiphon pisum. *Molecular biology and evolution*, 27(5) :979–87, May 2010.

[80] Benjamín Ortiz-Rivas, Stéphanie Jaubert-Possamai, Sylvie Tanguy, Jean-Pierre Gauthier, Denis Tagu, and Rispe Claude. Evolutionary study of duplications of the miRNA machinery in aphids associated with striking rate acceleration and changes in expression profiles. *BMC evolutionary biology*, 12(1) :216, January 2012.

[81] Fabrice Legeai, Guillaume Rizk, Thomas Walsh, Owain Edwards, Karl Gordon, Dominique Lavenier, Nathalie Leterme, Agnès Méreau, Jacques Nicolas, Denis Tagu, and Stéphanie Jaubert-Possamai. Bioinformatic prediction, deep sequencing of microRNAs and expression analysis during phenotypic plasticity in the pea aphid, Acyrthosiphon pisum. *BMC genomics*, 11(1) :281, January 2010.

[82] Lorenzo F Sempere, Edward B Dubrovsky, Veronica A Dubrovskaya, Edward M Berger, and Victor Ambros. The expression of the let-7 small regulatory RNA is controlled by ecdysone during metamorphosis in Drosophila melanogaster. *Developmental biology*, 244 :170–179, 2002.

[83] Aurore Gallot. *Mécanisme moléculaires de la différentiation du mode de reproduction du puceron du pois Acyrthosiphon pisum : étude des régulation transcriptionnelles et post-transcriptionnelles.* PhD thesis, Agrocampus Ouest - Université européenne de Bretagne, 2011.

[84] Bernhard Ganter and C. Wille, Rudolf/Translator-Franzke. *Formal Concept Analysis : Mathematical Foundations.* Springer-Verlag New York, Inc., December 1997.

[85] Sylvain Blachon, Ruggero G Pensa, Jérémy Besson, Céline Robardet, Jean-François Boulicaut, and Olivier Gandrillon. Clustering formal concepts to discover biologically relevant knowledge from gene expression data. *In silico biology*, 7(4-5) :467–83, January 2007.

[86] V Choi, Y Huang, V Lam, D Potter, R Laubenbacher, and K Duca. Using formal concept analysis for microarray data comparison. *Journal of bioinformatics and computational biology*, 6(1) :65–75, February 2008.

[87] Mehdi Kaytoue, Sergei O. Kuznetsov, Amedeo Napoli, and Sébastien Duplessis. Mining gene expression data with pattern structures in formal concept analysis. *Information Sciences*, 181(10) :1989–2001, May 2011.

[88] Anna Hristoskova, Veselka Boeva, and Elena Tsiporkova. A formal concept analysis approach to consensus clustering of multi-experiment expression data. *BMC bioinformatics*, 15(1) :151, January 2014.

[89] Jutta Gebert, Susanne Motameny, Ulrich Faigle, Christian V Forst, and Rainer Schrader. Identifying genes of gene regulatory networks using formal concept analysis. *Journal of computational biology : a journal of computational molecular cell biology*, 15(2) :185–94, March 2008.

[90] Johannes Wollbold, Reinhard Guthke, and Bernhard Ganter. Constructing a Knowledge Base for Gene Regulatory Dynamics by Formal Concept Analysis Methods. *K. Horimoto et al. (Eds.) : AB 2008, LNCS 5147. Springer, Heidelberg 2008, pp. 230-244*, July 2008.

[91] Chitta Baral. *Knowledge Representation, Reasoning, and Declarative Problem Solving*. Cambridge University Press, January 2003.

[92] Martin Gebser, Roland Kaminski, Benjamin Kaufmann, and Torsten Schaub. Answer Set Solving in Practice. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 6(3) :1–238, December 2012.

[93] Martin Gebser, Benjamin Kaufmann, Roland Kaminski, Max Ostrowski, Torsten Schaub, and Marius Schneider. Potassco : The Potsdam Answer Set Solving Collection. *AI Communications*, 24(2) :107–124, April 2011.

[94] MARTIN GEBSER, BENJAMIN KAUFMANN, and TORSTEN SCHAUB. Multi-threaded ASP solving with clasp. *Theory and Practice of Logic Programming*, 12(4-5) :525–545, September 2012.

[95] Jaime Huerta-Cepas, Marina Marcet-Houben, Miguel Pignatelli, Andrés Moya, and Toni Gabaldón. The pea aphid phylome : a complete catalogue of evolutionary histories and arthropod orthology and paralogy relationships for acyrthosiphon pisum genes. *Insect Molecular Biology*, 19(s2) :13–21, 2010.

[96] JIM HARDIE and A. D. LEES. The induction of normal and teratoid viviparae by a juvenile hormone and kinoprene in two species of aphids. *Physiological Entomology*, 10(1) :65–74, March 1985.

[97] Ana Conesa, Stefan Götz, Juan Miguel García-Gómez, Javier Terol, Manuel Talón, and Montserrat Robles. Blast2GO : a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics (Oxford, England)*, 21(18) :3674–6, September 2005.

[98] Stefan Götz, Juan Miguel García-Gómez, Javier Terol, Tim D Williams, Shivashankar H Nagaraj, María José Nueda, Montserrat Robles, Manuel Talón, Joaquín Dopazo, and Ana Conesa. High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic acids research*, 36(10) :3420–35, June 2008.

[99] M Ashburner, C A Ball, J A Blake, D Botstein, H Butler, J M Cherry, A P Davis, K Dolinski, S S Dwight, J T Eppig, M A Harris, D P Hill, L Issel-Tarver, A Kasarskis, S Lewis, J C Matese, J E Richardson, M Ringwald, G M Rubin, and G Sherlock. Gene ontology : tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, 25(1) :25–9, May 2000.

[100] S F Altschul, W Gish, W Miller, E W Myers, and D J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3) :403–10, October 1990.

[101] Nicola J Mulder, Rolf Apweiler, Teresa K Attwood, Amos Bairoch, Alex Bateman, David Binns, Peer Bork, Virginie Buillard, Lorenzo Cerutti, Richard Copley, Emmanuel Courcelle, Ujjwal Das, Louise Daugherty, Mark Dibley, Robert Finn, Wolfgang Fleischmann, Julian Gough, Daniel Haft, Nicolas Hulo, Sarah Hunter, Daniel Kahn, Alexander Kanapin, Anish Kejariwal, Alberto Labarga, Petra S Langendijk-Genevaux, David Lonsdale, Rodrigo Lopez, Ivica Letunic, Martin Madera, John Maslen, Craig McAnulla, Jennifer McDowall, Jaina Mistry, Alex Mitchell, Anastasia N Nikolskaya, Sandra Orchard, Christine Orengo, Robert Petryszak, Jeremy D Selengut, Christian J A Sigrist, Paul D Thomas, Franck Valentin, Derek Wilson, Cathy H Wu, and Corin Yeats. New developments in the InterPro database. *Nucleic acids research*, 35(Database issue) :D224–8, January 2007.

[102] E. M. Zdobnov and R. Apweiler. InterProScan - an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, 17(9) :847–848, September 2001.

[103] Marc R Friedländer, Sebastian D Mackowiak, Na Li, Wei Chen, and Nikolaus Rajewsky. mirdeep2 accurately identifies known and hundreds of novel microrna genes in seven animal clades. *Nucleic acids research*, 40(1) :37–52, 2012.

[104] Marc R Friedländer, Wei Chen, Catherine Adamidi, Jonas Maaskola, Ralf Einspanier, Signe Knespel, and Nikolaus Rajewsky. Discovering microRNAs from deep sequencing data using miRDeep. *Nature biotechnology*, 26(4) :407–15, April 2008.

[105] I. L. Hofacker. Vienna RNA secondary structure server. *Nucleic Acids Research*, 31(13) :3429–3431, July 2003.

[106] Eric Bonnet, Jan Wuyts, Pierre Rouzé, and Yves Van de Peer. Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. *Bioinformatics (Oxford, England)*, 20(17) :2911–7, November 2004.

[107] Aaron R. Quinlan and Ira M. Hall. Bedtools : a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6) :841–842, 2010.

[108] Kim D Pruitt, Garth R Brown, Susan M Hiatt, Françoise Thibaud-Nissen, Alexander Astashyn, Olga Ermolaeva, Catherine M Farrell, Jennifer Hart, Melissa J Landrum, Kelly M McGarvey, Michael R Murphy, Nuala A O'Leary, Shashikant Pujar, Bhanu Rajput, Sanjida H Rangwala, Lillian D Riddick, Andrei Shkeda, Hanzhen Sun, Pamela Tamez, Raymond E Tully, Craig Wallin, David Webb, Janet Weber, Wendy Wu, Michael DiCuccio, Paul Kitts, Donna R Maglott, Terence D Murphy, and James M Ostell. RefSeq : an update on mammalian reference sequences. *Nucleic acids research*, 42(Database issue) :D756–63, January 2014.

[109] Jaime Huerta-Cepas, Salvador Capella-Gutiérrez, Leszek P Pryszcz, Marina Marcet-Houben, and Toni Gabaldón. PhylomeDB v4 : zooming into the plurality of evolutionary histories of a genome. *Nucleic acids research*, 42(Database issue) :D897–902, January 2014.

[110] Susan E St Pierre, Laura Ponting, Raymund Stefancsik, and Peter McQuilton. FlyBase 102–advanced approaches to interrogating FlyBase. *Nucleic acids research*, 42(Database issue) :D780–8, January 2014.

[111] Michel J Weber. New human and mouse microRNA genes found by homology search. *The FEBS journal*, 272(1) :59–73, January 2005.

[112] Eugene Berezikov, Nicolas Robine, Anastasia Samsonova, Jakub O Westholm, Ammar Naqvi, Jui-Hung Hung, Katsutomo Okamura, Qi Dai, Diane Bortolamiol-Becet, Raquel Martin, Yongjun Zhao, Phillip D Zamore, Gregory J Hannon, Marco A Marra, Zhiping Weng, Norbert Perrimon, and Eric C Lai. Deep annotation of Drosophila melanogaster microRNAs yields insights into their processing, modification, and emergence. *Genome research*, 21(2) :203–15, February 2011.

[113] Josue Moura Romao, Weiwu Jin, Maolong He, Tim McAllister, and Le Luo Guan. MicroRNAs in bovine adipogenesis : genomic context, expression and function. *BMC genomics*, 15(1) :137, January 2014.

[114] Andrea Tanzer and Peter F Stadler. Molecular Evolution of a MicroRNA Cluster. *Journal of Molecular Biology*, 339(2) :327–335, 2004.

[115] Antonio Marco, Maria Ninova, and Sam Griffiths-Jones. Multiple products from microRNA transcripts. *Biochemical Society transactions*, 41(4) :850–4, August 2013.

[116] Michael J Axtell, Jakub O Westholm, and Eric C Lai. Vive la différence : biogenesis and evolution of microRNAs in plants and animals. *Genome biology*, 12(4) :221, January 2011.

[117] J D Thompson, D G Higgins, and T J Gibson. CLUSTAL W : improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research*, 22(22) :4673–80, November 1994.

[118] M A Larkin, G Blackshields, N P Brown, R Chenna, P A McGettigan, H McWilliam, F Valentin, I M Wallace, A Wilm, R Lopez, J D Thompson, T J Gibson, and D G Higgins. Clustal W and Clustal X version 2.0. *Bioinformatics (Oxford, England)*, 23(21) :2947–8, November 2007.

[119] E. Beitz. TeXshade : shading and labeling of multiple sequence alignments using LaTeX2e. *Bioinformatics*, 16(2) :135–139, February 2000.

[120] Paul Jaccard. Étude comparative de la distribution florale dans une portion des Alpes et du Jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37 :547–579, 1901.

[121] William Ritchie, Stephane Flamant, and John E J Rasko. Predicting microRNA targets and functions : traps for the unwary. *Nature methods*, 6(6) :397–8, June 2009.

[122] Yahong Xu, Shunwen Luo, Yang Liu, Jian Li, Yi Lu, Zhigang Jia, Qihua Zhao, Xiaoping Ma, Minghui Yang, Yue Zhao, Ping Chen, and Yu Guo. Integrated gene network analysis and text mining revealing PIK3R1 regulated by miR-127 in human bladder cancer. *European journal of medical research*, 18(1) :29, January 2013.

[123] Lian-Jie Lin, Yan Lin, Yu Jin, and Chang-Qing Zheng. Investigation of key microRNAs associated with hepatocellular carcinoma using small RNA-seq data. *Molecular biology reports*, March 2014.

[124] Mark D Robinson and Gordon K Smyth. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics (Oxford, England)*, 23(21) :2881–7, November 2007.

[125] Mark D Robinson and Gordon K Smyth. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics (Oxford, England)*, 9(2) :321–32, April 2008.

[126] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edgeR : a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, 26(1) :139–40, January 2010.

[127] Davis J McCarthy, Yunshun Chen, and Gordon K Smyth. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic acids research*, 40(10) :4288–97, May 2012.

[128] R Core Team. *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012. ISBN 3-900051-07-0.

[129] Mark D Robinson and Alicia Oshlack. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome biology*, 11(3) :R25, January 2010.

[130] Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1) :289 – 300, 1995.

[131] Fran Supek, Matko Bošnjak, Nives Škunca, and Tomislav Šmuc. Revigo summarizes and visualizes long lists of gene ontology terms. *PLoS ONE*, 6, 2011.

[132] Andreas Schlicker, Francisco S Domingues, Jörg Rahnenführer, and Thomas Lengauer. A new measure for functional similarity of gene products based on Gene Ontology. *BMC bioinformatics*, 7(1) :302, January 2006.

[133] Jianhao Jiang, Xie Ge, Zhiqian Li, Yueqiang Wang, Qisheng Song, David W Stanley, Anjiang Tan, and Yongping Huang. MicroRNA-281 regulates the expression of ecdysone receptor (EcR) isoform B in the silkworm, Bombyx mori. *Insect biochemistry and molecular biology*, 43(8) :692–700, August 2013.

[134] Valérie Hilgers, Natascha Bushati, and Stephen M. Cohen. Drosophila microRNAs 263a/b confer robustness during development by protecting nascent sense organs from apoptosis. *PLoS Biology*, 8, 2010.

[135] Aurelio A Teleman, Sushmita Maitra, and Stephen M Cohen. Drosophila lacking microRNA miR-278 are defective in energy homeostasis. *Genes & development*, 20 :417–422, 2006.

[136] Jenn-Yah Yu, Steven H Reynolds, Steve D Hatfield, Halyna R Shcherbata, Karin A Fischer, Ellen J Ward, Dang Long, Ye Ding, and Hannele Ruohola-Baker. Dicer-1-dependent Dacapo suppression acts downstream of Insulin receptor in regulating cell division of Drosophila germline stem cells. *Development (Cambridge, England)*, 136 :1497–1507, 2009.

[137] Sayaka Hori, Kumi Kaneko, Takeshi H. Saito, Hideaki Takeuchi, and Takeo Kubo. Expression of two microRNAs, ame-mir-276 and -1000, in the adult honeybee (Apis mellifera) brain. *Apidologie*, 42(1) :89–102, May 2011.

[138] Peizhang Xu, Stephanie Y. Vernooy, Ming Guo, and Bruce A. Hay. The Drosophila microRNA mir-14 suppresses cell death and is required for normal fat metabolism. *Current Biology*, 13 :790–795, 2003.

[139] Jishy Varghese and Stephen M Cohen. microRNA miR-14 acts to modulate a positive autoregulatory loop controlling steroid hormone signaling in Drosophila. *Genes & development*, 21 :2277–2282, 2007.

[140] Jishy Varghese, Sing Fee Lim, and Stephen M Cohen. Drosophila miR-14 regulates insulin production and metabolism through its target, sugarbabe. *Genes & development*, 24 :2748–2753, 2010.

[141] Regalla Kumarswamy and Sudhir Chandna. Inhibition of microRNA-14 contributes to actinomycin-D-induced apoptosis in the Sf9 insect cell line. *Cell biology international*, 34 :851–857, 2010.

[142] Nan Liu, Michael Landreh, Kajia Cao, Masashi Abe, Gert-Jan Hendriks, Jason R Kennerdell, Yongqing Zhu, Li-San Wang, and Nancy M Bonini. The microrna mir-34 modulates ageing and neurodegeneration in drosophila. *Nature*, 482(7386) :519–523, 2012.

[143] Chulan Kwon, Zhe Han, Eric N Olson, and Deepak Srivastava. MicroRNA1 influences cardiac differentiation in Drosophila and regulates Notch signaling. *Proceedings of the National Academy of Sciences of the United States of America*, 102 :18986–18991, 2005.

[144] Jérémy Besson, Céline Robardet, and Jean-Francois Boulicaut. Mining formal concepts with a bounded number of exceptions from transactional data. In *Post-Workshop Proceedings of the 3rd International Workshop on Knowledge Discovery in Inductive Databases KDID'04*, volume 3377, pages 33–45. Springer-Verlag LNCS 3377, 2004.

[145] Radim Belohlavek and Vilem Vychodil. Replacing full rectangles by dense rectangles : concept lattices and attribute implications. In *2006 IEEE International Conference on Information Reuse & Integration*, pages 117–122. IEEE, September 2006.

[146] Mikhail Klimushkin, Sergei Obiedkov, and Camille Roth. Approaches to the selection of relevant concepts in the case of noisy data. In *Formal Concept Analysis*, pages 255–266. Springer, 2010.

[147] J. Rissanen. Modeling by shortest data description. *Automatica*, 14(5) :465–471, September 1978.

[148] Chris Fraley and Adrian E Raftery. Model-Based Clustering, Discriminant Analysis, and Density Estimation. *Journal of the American Statistical Association*, 97(458) :611–631, June 2002.

[149] Chris Fraley, AE Raftery, and L Scrucca. Normal mixture modeling for model-based clustering, classification, and density estimation. *Department of Statistics, University of Washington, Available online at http ://cran. r-project. org/web/-packages/mclust/index. html. Accessed September*, 23 :2012, 2012.

[150] Gideon Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2) :461–464, March 1978.

[151] J Amilhastre, M.C Vilarem, and P Janssen. Complexity of minimum biclique cover and minimum biclique decomposition for bipartite domino-free graphs. *Discrete Applied Mathematics*, 86(2-3) :125–144, September 1998.

[152] Herbert Fleischner, Egbert Mujuni, Daniël Paulusma, and Stefan Szeider. Covering graphs with few complete bipartite subgraphs. *Theoretical Computer Science*, 410(21-23) :2045–2053, May 2009.

[153] Yun Zhang, Charles A Phillips, Gary L Rogers, Erich J Baker, Elissa J Chesler, and Michael A Langston. On finding bicliques in bipartite graphs : a novel algorithm and its application to the integration of diverse biological data types. *BMC bioinformatics*, 15(1) :110, January 2014.

[154] Loïc Royer, Matthias Reimann, Bill Andreopoulos, and Michael Schroeder. Unraveling protein networks with power graph analysis. *PLoS computational biology*, 4(7) :e1000108, January 2008.

[155] Martina Maisel, Hans-Jörg Habisch, Loïc Royer, Alexander Herr, Javorina Milosevic, Andreas Hermann, Stefan Liebau, Rolf Brenner, Johannes Schwarz, Michael Schroeder, and Alexander Storch. Genome-wide expression profiling and functional network analysis upon neuroectodermal conversion of human mesenchymal stem cells suggest HIF-1 and miR-124a as important regulators. *Experimental cell research*, 316(17) :2760–78, October 2010.

[156] Li Li, David J Ruau, Chirag J Patel, Susan C Weber, Rong Chen, Nicholas P Tatonetti, Joel T Dudley, and Atul J Butte. Disease risk factors identified through shared genetic architecture and electronic medical records. *Science translational medicine*, 6(234) :234ra57, April 2014.

[157] George Tsatsaronis, Matthias Reimann, Iraklis Varlamis, Orestis Gkorgkas, and Kjetil Nø rvåg. Efficient community detection using power graph analysis. In *Proceedings of the 9th workshop on Large-scale and distributed informational retrieval - LSDS-IR '11*, page 21, New York, New York, USA, October 2011. ACM Press.

[158] George Tsatsaronis, Iraklis Varlamis, Sunna Torge, Matthias Reimann, Kjetil Nørvåg, Michael Schroeder, and Matthias Zschunke. How to become a group leader ? or modeling author types based on graph mining. In *Research and Advanced Technology for Digital Libraries*, pages 15–26. Springer, 2011.

[159] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25) :14863–14868, December 1998.

[160] Thomas H Cormen, Charles E Leiserson, Ronald L Rivest, Clifford Stein, et al. *Introduction to algorithms*, volume 2. MIT press Cambridge, 2001.

[161] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape : a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13 :2498–2504, 2003.

[162] Emden R. Gansner and Stephen C. North. An open graph visualization system and its applications to software engineering. *SOFTWARE - PRACTICE AND EXPERIENCE*, 30(11) :1203–1233, 2000.

[163] Carsten Skurk, Henrike Maatz, Hyo-Soo Kim, Jiang Yang, Md Ruhul Abid, William C Aird, and Kenneth Walsh. The Akt-regulated forkhead transcription factor FOXO3a controls endothelial cell viability through modulation of the caspase-8 inhibitor FLIP. *The Journal of biological chemistry*, 279(2) :1513–25, January 2004.

[164] Thomas C Brionne, Ina Tesseur, Eliezer Masliah, and Tony Wyss-Coray. Loss of TGF-beta 1 leads to increased neuronal cell death and microgliosis in mouse brain. *Neuron*, 40 :1133–1145, 2003.

[165] Kiyomi Yoshida, Masashi Yamada, Chika Nishio, Akio Konishi, and Hiroshi Hatanaka. SNRK, a member of the SNF1 family, is related to low K+-induced apoptosis of cultured rat cerebellar granule neurons. *Brain Research*, 873(2) :274–282, August 2000.

[166] Ashish Thapliyal, Rashmi Verma, and Navin Kumar. Small G Proteins Dexras1 and RHES and Their Role in Pathophysiological Processes. *International journal of cell biology*, 2014 :308535, January 2014.

[167] Sushmita Roy, Jason Ernst, Peter V Kharchenko, Pouya Kheradpour, Nicolas Negre, Matthew L Eaton, Jane M Landolin, Christopher a Bristow, Lijia Ma, Michael F Lin, Stefan Washietl, Bradley I Arshinoff, Ferhat Ay, Patrick E Meyer, Nicolas Robine, Nicole L Washington, Luisa Di Stefano, Eugene Berezikov, Christopher D Brown, Rogerio Candeias, Joseph W Carlson, Adrian Carr, Irwin Jungreis, Daniel Marbach, Rachel Sealfon, Michael Y Tolstorukov, Sebastian Will, Artyom a Alekseyenko, Carlo Artieri, Benjamin W Booth, Angela N Brooks, Qi Dai, Carrie a Davis, Michael O Duff, Xin Feng, Andrey a Gorchakov, Tingting Gu, Jorja G Henikoff, Philipp Kapranov, Renhua Li, Heather K MacAlpine, John Malone, Aki Minoda, Jared Nordman, Katsutomo Okamura, Marc Perry, Sara K Powell, Nicole C Riddle, Akiko Sakai, Anastasia Samsonova, Jeremy E Sandler, Yuri B Schwartz, Noa Sher, Rebecca Spokony, David Sturgill, Marijke van Baren, Kenneth H Wan, Li Yang, Charles Yu, Elise Feingold, Peter Good, Mark Guyer, Rebecca Lowdon, Kami Ahmad, Justen Andrews, Bonnie Berger, Steven E Brenner, Michael R Brent, Lucy Cherbas, Sarah C R Elgin, Thomas R Gingeras, Robert Grossman, Roger a Hoskins, Thomas C Kaufman, William Kent, Mitzi I Kuroda, Terry Orr-Weaver, Norbert Perrimon, Vincenzo Pirrotta, James W Posakony, Bing Ren, Steven Russell, Peter Cherbas, Brenton R Graveley, Suzanna Lewis, Gos Micklem, Brian Oliver, Peter J Park, Susan E Celniker, Steven Henikoff, Gary H Karpen, Eric C Lai, David M MacAlpine, Lincoln D Stein, Kevin P White, and Manolis Kellis. Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science (New York, N.Y.)*, 330(6012) :1787–97, December 2010.

[168] Leah R Sabin, Rui Zhou, Joshua J Gruber, Nina Lukinova, Shelly Bambina, Allison Berman, Chi-Kong Lau, Craig B Thompson, and Sara Cherry. Ars2 regulates both miRNA- and siRNA- dependent silencing and suppresses RNA virus infection in Drosophila. *Cell*, 138(2) :340–51, July 2009.

[169] Joshua J Gruber, D Steven Zatechka, Leah R Sabin, Jeongsik Yong, Julian J Lum, Mei Kong, Wei-Xing Zong, Zhenxi Zhang, Chi-Kong Lau, Jason Rawlings, Sara Cherry, James N Ihle, Gideon Dreyfuss, and Craig B Thompson. Ars2 links the nuclear cap-binding complex to RNA interference and cell proliferation. *Cell*, 138(2) :328–39, July 2009.

[170] E H Baehrecke. who encodes a KH RNA binding protein that functions in muscle development. *Development (Cambridge, England)*, 124(7) :1323–32, April 1997.

[171] J Huerta-Cepas, M Marcet-Houben, M Pignatelli, A Moya, and T Gabaldón. The pea aphid phylome : a complete catalogue of evolutionary histories and arthropod orthology and paralogy relationships for Acyrthosiphon pisum genes. *Insect molecular biology*, 19 Suppl 2 :13–21, March 2010.

[172] John K Colbourne, Michael E Pfrender, Donald Gilbert, W Kelley Thomas, Abraham Tucker, Todd H Oakley, Shinichi Tokishita, Andrea Aerts, Georg J Arnold, Malay Kumar Basu, Darren J Bauer, Carla E Cáceres, Liran Carmel, Claudio Casola, Jeong-Hyeon Choi, John C Detter, Qunfeng Dong, Serge Dusheyko, Brian D

Eads, Thomas Fröhlich, Kerry A Geiler-Samerotte, Daniel Gerlach, Phil Hatcher, Sanjuro Jogdeo, Jeroen Krijgsveld, Evgenia V Kriventseva, Dietmar Kültz, Christian Laforsch, Erika Lindquist, Jacqueline Lopez, J Robert Manak, Jean Muller, Jasmyn Pangilinan, Rupali P Patwardhan, Samuel Pitluck, Ellen J Pritham, Andreas Rechtsteiner, Mina Rho, Igor B Rogozin, Onur Sakarya, Asaf Salamov, Sarah Schaack, Harris Shapiro, Yasuhiro Shiga, Courtney Skalitzky, Zachary Smith, Alexander Souvorov, Way Sung, Zuojian Tang, Dai Tsuchiya, Hank Tu, Harmjan Vos, Mei Wang, Yuri I Wolf, Hideo Yamagata, Takuji Yamada, Yuzhen Ye, Joseph R Shaw, Justen Andrews, Teresa J Crease, Haixu Tang, Susan M Lucas, Hugh M Robertson, Peer Bork, Eugene V Koonin, Evgeny M Zdobnov, Igor V Grigoriev, Michael Lynch, and Jeffrey L Boore. The ecoresponsive genome of Daphnia pulex. *Science (New York, N.Y.)*, 331(6017) :555–61, February 2011.

[173] Jean-Christophe Simon, Michael E Pfrender, Ralph Tollrian, Denis Tagu, and John K Colbourne. Genomics of environmentally induced phenotypes in 2 extremely plastic arthropods. *The Journal of heredity*, 102(5) :512–25, January 2011.

[174] Geng Tian, XuYang Yin, Hong Luo, XiaoHong Xu, Lars Bolund, XiuQing Zhang, Shang-Quang Gan, and Ning Li. Sequencing bias : comparison of different protocols of microRNA library construction. *BMC biotechnology*, 10(1) :64, January 2010.

[175] J J Sekelsky, S J Newfeld, L A Raftery, E H Chartoff, and W M Gelbart. Genetic characterization and cloning of mothers against dpp, a gene required for decapentaplegic function in Drosophila melanogaster. *Genetics*, 139(3) :1347–58, March 1995.

[176] Emily Clough, Woongjoon Moon, Shengxian Wang, Kathleen Smith, and Tulle Hazelrigg. Histone methylation is required for oogenesis in Drosophila. *Development (Cambridge, England)*, 134(1) :157–65, January 2007.

[177] Mohamed Rouane-Hacene, Marianne Huchard, Amedeo Napoli, and Petko Valtchev. Relational concept analysis : mining concept lattices from multi-relational data. *Annals of Mathematics and Artificial Intelligence*, 67(1) :81–108, March 2013.

[178] John L Rinn and Howard Y Chang. Genome regulation by long noncoding RNAs. *Annual review of biochemistry*, 81 :145–66, January 2012.

[179] Mitchell Guttman and John L Rinn. Modular regulatory principles of large noncoding RNAs. *Nature*, 482(7385) :339–46, February 2012.

[180] Alessandro Fatica and Irene Bozzoni. Long non-coding RNAs : new players in cell differentiation and development. *Nature reviews. Genetics*, 15(1) :7–21, January 2014.

Annexes

Edge Selection in a Noisy Graph by Concept Analysis: Application to a Genomic Network

Valentin Wucher[1,2], Denis Tagu[1], and Jacques Nicolas[2]

[1] INRA, UMR 1349 IGEPP, Le Rheu, 35653, France
valentin.wucher@rennes.inra.fr; denis.tagu@rennes.inra.fr
[2] IRISA-INRIA, Campus de Beaulieu, 35042 Rennes cedex, France
jacques.nicolas@inria.fr

Abstract. MicroRNAs (miRNAs) are small RNA molecules that bind messenger RNAs (mRNAs) to silence their expression. Understanding this regulation mechanism requires the study of the miRNA/mRNA interaction network. State of the art methods for predicting interactions lead to a high level of false positive: the interaction score distribution may be roughly described as a mixture of two overlapping Gaussian laws that need to be discriminated with a threshold. In order to further improve the discrimination between true and false interactions, we present a method that considers the structure of the underlying graph. We assume that the graph is formed on a relatively simple structure of formal concepts (associated to regulation modules in the regulation mechanism). Specifically, the formal context topology of true edges is assumed to be less complex than in the case of a noisy graph including spurious interactions or missing interactions. Our approach consists thus in selecting edges below an edge score threshold and applying a repair process on the graph, adding or deleting edges to decrease the global concept complexity. To validate our hypothesis and method, we have extracted parameters from a real biological miRNA/mRNA network and used them to build random networks with fixed concept topology and true/false interaction ratio. Each repaired network can be evaluated with a score balancing the number of edge changes and the conceptual adequacy in the spirit of the minimum description length principle.

1 Introduction

MicroRNAs (miRNAs) are small RNA molecules that bind to and regulate the flow of messenger RNAs (mRNAs). They have a sequence of 6 nucleotides that bind to a complementary sequence, the binding site, of the target mRNA. Bound miRNAs repress the expression of their target mRNAs.

The interaction network created by miRNAs/mRNAs interactions is by definition a bipartite graph between miRNA nodes and mRNA nodes. Several bioinformatics methods can predict miRNAs/mRNAs interactions. The current state of the art offers only methods having a high level of false positive predictions (Chil et al. (2009), Reyes-Herrera et al. (2011)). Even with

scoring functions and a threshold, it is still hard to discriminate between true and false predictions.

Based on the biological function of miRNAs, i.e. repressing mRNAs translation, and their implication in many biological processes (Janga and Vallabhanenis (2011)), authors have provided some evidence that miRNAs combine to regulate functional modules, i.e. clusters of mRNAs sharing similar functions (Bryan et al. (2014) and Enright et al. (2005)). This assumption is compatible with the observations of similar complexes for another major regulation actor, transcription factors. Thus true interactions could be distinguished in principle from false one on the basis of functional clusters (modules), i.e. set of miRNAs that regulate mRNAs with the same function.

Once a score threshold has been set, we intend to improve edge selection by detecting false negatives and false positives by taking into account the previous assumption in the framework of formal concept analysis.

2 Definition of formal concept analysis

This section briefly recalls some notions of formal concept analysis as defined by Ganter and Wille (1999) and Klimushkin et al. (2010).

A *formal context* is a triplet $\mathbb{K} = (G, M, I)$ where G is the set of objects, M the set of attributes and $I \subseteq G \times M$ is a binary relation between objects and attributes. The operator $(.)'$ is defined on $\mathbb{K}$ for $A \subseteq G$ and $B \subseteq M$ as: $A' = \{m \in M | \forall g \in A : gIm\}$ and $B' = \{g \in G | \forall m \in B : gIm\}$. A' is the set of common attributes to all objects in A and B' the set of common objects to all attributes in B.

A *formal concept* is a pair (A, B) defined on $\mathbb{K}$ with $A \subseteq G$ and $B \subseteq M$ where $A = B'$ and $B = A'$. Concept ordering can be based on set inclusion: For all formal concepts (A, B) and (C, D), let $(A, B) \leq (C, D)$ if $A \subseteq C$. If $(A, B) \leq (C, D)$ and there is no formal concept (E, F) such that $(A, B) < (E, F) < (C, D)$ then we write $(A, B) \prec (C, D)$.

The relation $<$ generates a *concept lattice* structure $\underline{\mathfrak{B}}(\mathbb{K})$ on context $\mathbb{K}$. The order $\prec$ generates the edges in the covering graph of $\underline{\mathfrak{B}}(\mathbb{K})$.

	a_1	a_2	a_3	a_4
o_1	1	1		
o_2	1	1		
o_3	1	1		
o_4			1	1
o_5			1	1

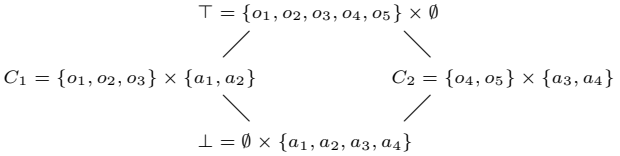


Fig. 1. A formal context $\mathbb{K}_{\text{ex}}$ (left) with $G_{\text{ex}} = \{o_{1..5}\}$ the set of objects and $M_{\text{ex}} = \{a_{1..4}\}$ the set of attributes and the associated concept lattice $\underline{\mathfrak{B}}(\mathbb{K}_{\text{ex}})$ (right).

Figure 1 gives a small example of formal context and the associated concept lattice. It contains four formal concepts, namely C_1, C_2, the top concept $\top$ and the bottom concept $\bot$.

3 The effect of noise on formal concept analysis

Formal concept analysis is a powerful method for binary data analysis because it extracts every complete group of related elements, i.e. such that every element from one set is related to every element in the second set. This advantage become a drawback in case of noisy data, because of its sensitivity to the presence of each relation.

Studies have already been conducted on fault-tolerant or approximated concepts analysis (Besson et al. (2004), Belohlavek and Vychodil (2006), Blachon et al. (2007)). It consists mostly in retrieving dense rectangles of 1 in a binary matrix: a concept is indeed a submatrix filled with 1 values, up to line and column reordering. The constraint of requiring a complete set of 1 may be released by an optimisation constraint requiring a maximal number of 1. Very few works exist aiming at retrieving original concepts from noisy formal concepts. One of the most advanced study in this domain is due to Klimushkin et al. (2010), which showed that formal concepts can be recovered from a formal context including false relations and between 300 to 400 objects and 4 to 12 attributes. They used three statistical values on concepts to find the original concepts and concept lattice.

The next subsection introduces a toy example of noisy context in order to illustrate the effect of noise on the associated concept lattice. A more formal characterization of this effect is provided in a subsection.

3.1 Example of noise effect

In the context $\mathbb{K}_{\text{noise}}$ (Figure 2), one spurious relation (o_5, a_2) has been added compared with Figure 1 and a dissimilarity score is available for every relation. By setting a threshold of -0.2 and keeping every relation below this threshold, (o_3, a_2), an original relation, is discarded while (o_5, a_2), a spurious relation, is kept. There are now 7 concepts, 3 more than in $\mathbb{K}_{\text{ex}}$ (see Figure 2). The deletion of element (o_3, a_2) has split concept C_1 into two concepts C_1' and C_1''. Concept C_2 still exists in $\mathbb{K}_{\text{noise}}$, renamed C_2' in the figure. Two new concepts, C_3 and C_4 have been created due to the addition of (o_5, a_2).

	a_1	a_2	a_3	a_4
o_1	-0.3	-0.25		
o_2	-0.5	-0.45		
o_3	-0.6	-0.1		
o_4			-0.4	-0.2
o_5		<u>-0.3</u>	-0.28	-0.41

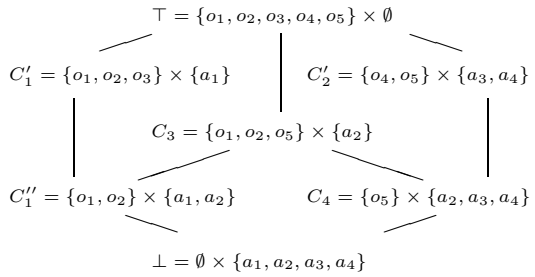


Fig. 2. Formal context $\mathbb{K}_{\text{noise}}$ (left) with scores (spurious relation underlined) and the associated concept lattice $\underline{\mathfrak{B}}(\mathbb{K}_{\text{noise}})$ (right) obtained at threshold -0.2.

3.2 Spurious relations

To better understand the local effect of spurious relations on a concept, we need to discriminate two types of relations: the original relations, $I^o \subseteq G \times M$ and the spurious relations, $I^s \subseteq G \times M$ with $I^o \cap I^s = \emptyset$. These two types of relations involve three types of contexts: the original context with no spurious relation $\mathbb{K}^o = (G, M, I)$, the context containing only the spurious relations $\mathbb{K}^s = (G, M, I^s)$ and the context with all the relations $\mathbb{K}^{os} = (G, M, (I \cup I^s))$. They generate three types of formal concepts, the set of original concepts $\mathfrak{C}^o$ defined on $\mathbb{K}^o$, the set of spurious concepts $\mathfrak{C}^s$ defined on $\mathbb{K}^s$ and the set of concepts $\mathfrak{C}^{os}$ defined on $\mathbb{K}^{os}$. The construction of $\mathfrak{C}^{os}$ from $\mathfrak{C}^o$ and $\mathfrak{C}^s$ depends on the contribution of each concept pair in $\mathfrak{C}^o \times \mathfrak{C}^s$.

Consider $C^o = (A^o, B^o) \in \mathfrak{C}^o$ and $C^s = (A^s, B^s) \in \mathfrak{C}^s$. Since I^o and I^s are exclusive, the concepts in $\mathfrak{C}^o$ and $\mathfrak{C}^s$ need to be disjoint. It means that either $A^o \cap A^s = \emptyset$ or $B^o \cap B^s = \emptyset$.

Assume with no lack of generality that $A^o \cap A^s \neq \emptyset$ and $B^o \cap B^s = \emptyset$. Then a new concept $C^{os} = (A^{os}, B^{os})$ may be created with $A^{os} = A^o \cap A^s$ and $B^{os} = B^o \cup B^s$. Note that if $A^s \subseteq A^o$ (resp. $A^o \subseteq A^s$), then C^s (resp. C^o) is not maximal in $\mathbb{K}^{os}$ since it is included in C^{os}.

Formally the contribution of two disjoint concepts to the set of extended concepts $\mathfrak{C}^{os}$ can be defined through the application of an inclusion operator:

Definition 1. *The inclusion operator $i(.,.)$ is defined for a pair of disjoint concepts $(C^i, C^j) = ((A^i, B^i), (A^j, B^j))$ as $i(C^i, C^j) = \mathfrak{C}^{i \cup j}$ where $\mathfrak{C}^{i \cup j}$ is the set of concepts obtained on relation $\{(A^i \times B^i) \cup (A^j \times B^j)\}$.*

The various types of results deriving from i application depending on the intersection between object or attribute sets are listed below:

$$
\begin{aligned}
i(C^i, C^j) &= \{C^i, C^j\} && \text{if } A^i \cap A^j = B^i \cap B^j = \emptyset; && (1)\\
&= \{(A^i \cup A^j, B^i \cup B^j)\} && \text{if } A^i = A^j \text{ or } B^i = B^j; && (2)\\
&= \{C^j, (A^i \cup A^j, B^i \cup B^j)\} && \text{if } A^i \subset A^j \text{ or } B^i \subset B^j; && (3)\\
&= \{C^i, C^j, (A^i \cap A^j, B^i \cup B^j)\} && \text{if } A^i \cap A^j \not\subseteq \{\emptyset, A^i, A^j\}; && (4)\\
&= \{C^i, C^j, (A^i \cup A^j, B^i \cap B^j)\} && \text{if } B^i \cap B^j \not\subseteq \{\emptyset, B^i, B^j\}. && (5)
\end{aligned}
$$

$\mathfrak{C}^{os}$ can be defined using a fixpoint characterization: $\mathfrak{C}^{os}$ is the smallest set of concepts that cover the concepts of $\mathfrak{C}^s$ and $\mathfrak{C}^o$ and is closed under i. Concepts from $\mathfrak{C}^s$ and $\mathfrak{C}^o$ and concepts generated by operator i belong to $\mathfrak{C}^{os}$ if they are not covered by other concepts from $\mathfrak{C}^{os}$ as described above.

3.3 Missing relations

As for spurious relations, one can proceed by distinguishing two types of relations: the original relations, $I^o \subseteq G \times M$ and the missing relations, $I^m \subseteq I^o$. They imply three types of contexts: the original context without missing relations $\mathbb{K}^o = (G, M, I^o)$, the context containing only the missing relations $\mathbb{K}^m = (G, M, I^m)$ and the context with all except the missing relations $\mathbb{K}^{om} =$

$(G, M, (I^o \setminus I^m))$. These contexts entail three types of formal concepts, the set of original concepts $\mathfrak{C}^o$ defined on $\mathbb{K}^o$, the set of missing concepts $\mathfrak{C}^m$ defined on $\mathbb{K}^m$ and the set of concepts $\mathfrak{C}^{om}$ defined on $\mathbb{K}^{om}$. As for spurious relations, the objective is to describe how the sets $\mathfrak{C}^o$ and $\mathfrak{C}^m$ are combining in $\mathfrak{C}^{om}$. We first describe the general case where the relations of a concept in $\mathfrak{C}^m$ are overlapping those of a concept in $\mathfrak{C}^o$.

Consider $C^o = (A^o, B^o) \in \mathfrak{C}^o$ and $C^m = (A^m, B^m) \in \mathfrak{C}^m$, if $A^o \cap A^m \neq \emptyset$ and $B^o \cap B^m \neq \emptyset$, then the concept C^o cannot be in $\mathfrak{C}^{om}$ since it includes missing relations $A^m \times B^m$. Instead, two new concepts will be created in $\mathfrak{C}^{om}$, $C_1^{om} = (A^o, B^o \setminus B^m)$ and $C_2^{om} = (A^o \setminus A^m, B^o)$. Note that if $A^o \subseteq A^m$ (resp. $B^o \subseteq B^m$), then only the concept C_1^{om} is created (resp. C_2^{om}).

Formally the contribution of two overlapping concepts to the set of restricted concepts $\mathfrak{C}^{om}$ can be defined through the application of an exclusion operator:

Definition 2. *The exclusion operator $e(.,.)$ is defined for a pair of overlapping concepts $(C^i, C^j) = ((A^i, B^i), (A^j, B^j))$ as $e(C^i, C^j) = \mathfrak{C}^{j\setminus i}$ where $\mathfrak{C}^{j\setminus i}$ is the set of concepts obtained on relation $\{(A^j \times B^j) \setminus (A^i \times B^i)\}$.*

The various types of results deriving from e application depending on the intersection between object and attribute sets are listed below:

$$
\begin{aligned}
e(C^i, C^j) &= C^j && \text{if } A^j \cap A^i \text{ or } B^j \cap B^i = \emptyset; && (6)\\
&= \{(A^j, B^j \setminus B^i), (A^j \setminus A^i, B^j)\} && \text{if } A^j \cap A^i \neq \emptyset, B^j \cap B^i \neq \emptyset; && (7)\\
&= \{(A^j, B^j \setminus B^i)\} && \text{if } A^j \subseteq A^i, B^j \not\subseteq B^i; && (8)\\
&= \{(A^j \setminus A^i, B^j)\} && \text{if } A^j \not\subseteq A^i, B^j \subseteq B^i; && (9)\\
&= \emptyset && \text{if } A^j \subseteq A^i, B^j \subseteq B^i. && (10)
\end{aligned}
$$

$\mathfrak{C}^{om}$ can be defined using a fixpoint characterization: $\mathfrak{C}^{om}$ is the largest set of concepts which are included in the concepts of $\mathfrak{C}^o$ and is closed under e. Concepts from $\mathfrak{C}^o$ and concepts generated by operator e belong to $\mathfrak{C}^{om}$ if they do not contain a relation of I^m as described above.

3.4 Managing the noise

The previous study points out that the number of concepts will increase depending on the type of noisy relations (spurious or missing) and the number of purely noisy concepts except for equations (1) and (6) where no new concepts are created. For spurious relations, the number of new concepts in $\mathfrak{C}^{os}$ is bounded by the number n_s of disjoint concepts $C_j^s \in \mathfrak{C}^s$ with only one set that intersect with C and is bounded by n_s. For missing relations, the number of new concepts $C_i^{om} \in \mathfrak{C}^{om}$ locally created from a concept C depends on the number n_m of concepts $C_j^m \in \mathfrak{C}^m$ that overlap with C and is bounded by 2^{n_m}. Overall, the evolution of the number of new concepts is linear when adding spurious concepts and exponential when deleting missing concepts. To

repair the context $\mathbb{K}^{osm}$ (the context with $I^{osm} = ((I^o \cup I^s) \setminus I^m)$) in order to retrieve $\mathbb{K}^o$, we need to define new operations that reverse the effect of operators i and e. These operations take advantage of the fact that most of the time, concepts resulting from the application of i or e are connected in the concept lattice by a direct relation or a sibling relation.

Concerning operator i, in equation (3) the two result concepts are ordered by $\prec$ in the concept lattice. As for equations (4) and (5), the new concept is the direct precursor or the direct successor of C^i and C^j in the concept lattice. For operator e, in equation (7) the result concepts are ordered by $\prec$. The two sets A^j and B^j of the original concept can be easily recovered by crossing the noisy concepts.

4 Repair process

4.1 Definition of repair operations

We have introduced two operations *delete* and *add* resp. defined from operators i and e, which select then suppress or insert relations based on concept lattice analysis. We assume in the following that these operations act on a pair of concepts (X, Y) with $X = (A, B)$ and $Y = (C, D)$.

Two types of selected (X, Y) pair selection exist, link pair if $X \prec Y$ and sibling pair, $X \approx Y$, if $\exists Z | X \prec Z$ and $Y \prec Z$. The following operations apply on these pairs:

$$\forall\ (X,Y) \mid X \prec Y \text{ or } Y \prec X \text{ or } X \approx Y :$$
$$delete(X,Y) : \mathfrak{C} := \mathfrak{C} - Y; \qquad Noise := Noise \cup Y \setminus X;$$
$$\forall\ (X,Y) \mid X \prec Y \text{ or } Y \prec X :$$
$$add(X,Y) : \mathfrak{C} := \mathfrak{C} - X - Y + (C,B); \qquad Noise := Noise \cup (C \setminus A) \times (B \setminus D);$$

where $\mathfrak{C}$, initially the set of observed concepts, is the resulting set of concepts and $Noise$ is the set of spurious or missing interactions.

In Figure 2, $(C_2', C_4)_l$ and $delete(C_2', C_4) = delete((\{o_4, o_5\}, \{a_3, a_4\}), (\{o_5\}, \{a_2, a_3, a_4\}))$ results in deleting a spurious relation: $\mathfrak{C} := \mathfrak{C} - C_4$ and $Noise := Noise \cup \{(o_5, a_2)\}$. The same way, $(C_1'', C_1')_l$ and $add(C_1'', C_1') = add((\{o_1, o_2\}, \{a_1, a_2\}), (\{o_1, o_2, o_3\}, \{a_1\}))$ results in adding a missing relation: $\mathfrak{C} := \mathfrak{C} - C_1'' - C_1' + (\{o_1, o_2, o_3\}, \{a_1, a_2\})$ and $Noise := Noise \cup \{(o_3, a_2)\}$

The whole repair process consist of the simultaneous application of a set of *delete*/*add* operations on a subset of pairs extracted from the initially observed set of concepts. The space of admissible pairs is naturally constrained: once a concept has been chosen for deletion for instance, it cannot be used for an *add* operation in another selection. This leads to a space of different subsets of concepts, induced by different repair alternatives. These alternatives have to be scored in order to keep the best one.

4.2 Minimum description length optimization

In the spirit of the minimum description length principle, each set of concepts $\mathfrak{C}$ resulting from the application of *delete*/*add* operations on a subset of concept pairs $\mathfrak{C}$ gets a score defined as:

$$score(\mathfrak{C}) = \sum_{(A,B)\in\mathfrak{C}} (|A| + |B|) + \alpha\ card(Noise);$$

where α is an integer parameter set by default to 1. This score is minimized over all possible applications of *delete* and *add* operations on all admissible concept subsets $\mathfrak{C}$.

5 Experiments on simulated noisy data

We have generated several random contexts with a fixed number of objects and attributes (from 20 to 40) to test our method for the detection of spurious interactions. For each context, 5 sets of interactions have been created, corresponding to 5 cross-products of a random number of objects and a random number of attributes each following a normal distribution (mean 5, standard deviation 2). The original concepts are obtained on these sets of interactions. The noisy concepts are obtained by adding a uniform noise with a fixed probability for each cell to be changed. For each set of parameters (number of objects/attributes, noise level and weight α) 1,000 random contexts have been tried and the average ratio of original and spurious relations deleted has been computed. Results for the delete operation are shown Table 1.

Table 1. Mean and standard deviation (sd) on simulated noisy data of the percentage of original and spurious relations deleted by the repair process.

objects	attributes	noise	α	original (%)		spurious (%)	
				mean	sd	mean	sd
20	20	0.05	1	3.5	4.2	43.1	21.3
			2	3.3	4.1	43	21.3
			3	3.1	4	42.2	21.3
40	40	0.01	1	3.5	4.3	60.4	16
			2	2.3	3.1	60.6	15.9
			3	1.5	2.4	58.9	16.2
		0.05	1	0.4	1.5	29	11.2
			2	0.4	1.3	27.8	10.4
			3	0.3	0.1	24.2	9.1
		0.1	1	0.1	1.4	0.8	3.9
			2	0.1	1.4	0.8	3.9
			3	0	0	0.4	0.9
60	60	0.05	1	0.1	0.3	17.4	8.4
			2	0	0.3	15.7	6.8
			3	0	0.3	13.1	6.1

For all experiments in Table 1, the percentage of original and spurious relations deleted decreases when α increases. This observation is coherent with the defined score since α represents the relative minor importance of the number of deleted relations with respect to the description length of repaired concepts. In all cases, very few original concepts were affected by deletions. Half of spurious relations are detected for contexts that do not exceed 40 objects and 40 attributes and this rate decreases in line with context size increase and noise level increase. These results seemed sufficient for real data management with a relatively stringent selection of interactions (a limited level of noise) and we have further experimented with more realistic data.

6 miRNAs/mRNAs interaction graph

The pea aphid (*Acyrthosiphon pisum*) is a crop pest that is a model for the study of phenotypic plasticity (The Intern. Aphid Genomics Consortium (2010)). During the warm seasons viviparous parthenogenetic females are produced whereas during autumn, sexual males and oviparous sexual females are produced. In order to understand the differences in the regulation between sexual and asexual embryogenesis, kinetic data for mRNA and miRNA expression have been collected in both contexts (Gallot (2012)). From these data, we have extracted 43 miRNAs and 2,033 mRNAs of interest exhibiting kinetics differences in the two embryogenesis.

To predict miRNAs/mRNAs interactions, we used TargetScan v5 (Grimson et al. (2007)). TargetScan provides for each prediction a dissimilarity score, i.e. the lower the score, the stronger the interaction. This resulted in a scored interaction graph with 6,763 interactions between 41 miRNAs and 1,479 mRNAs. The prediction score distribution is shown Figure 3.

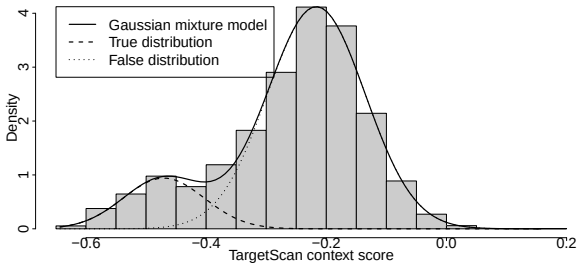


Fig. 3. TargetScan context score distribution (solid line: complete; dashed line: true prediction distribution; dotted line: spurious interactions).

The total distribution can be seen as a Gaussian mixture model (GMM, solid line) divided into two Gaussian curves, one centered around low values (dashed line) and the other centered around high values (spurious interactions, dotted lines). These curves are in agreement with the literature (Chil et al. (2009), Reyes-Herrera and Ficarra (2011)), which reveals a high false positive rate in the prediction methods. The TargetScan prediction forms thus a bipartite graph with a high level of noise, something that does not allow to directly apply our method. Fortunately, it is possible to select the most interesting interactions by choosing a relatively stringent score threshold. Every true interactions above this threshold will be missing and all spurious interactions below this threshold will be retained.

7 Experiments on simulated biological data

Since no miRNAs/mRNAs interactions dataset exist where actual interactions are known, we needed to simulate the interaction graphs in a controlled way

to test our method. The scores were simulated by fitting a GMM to the data (solid line in Figure 3). The degree of miRNAs and mRNAs vertices were determined using score dependent degree distributions for true interactions and spurious interactions.

We have generated 1,000 random miRNA/mRNA interactions graphs, keeping the number of miRNAs, mRNAs and interactions in real data: 41 miRNAs, 1,479 mRNAs and 6,763 interactions. Two thresholds have been tested, -0.3 and a more stringent one of -0.35, to restrict the number of spurious interactions while keeping a high number of true interactions (see Figure 3). A threshold defines the set of original interactions (true interactions below the threshold), spurious interactions (false interactions below the threshold) and missing interactions (true interactions above the threshold).

For each graph and each threshold, the concept lattice has been computed and the ratio of deleted original and spurious relations has been obtained for $\alpha = 1$. Results for the *delete* operation are shown in Table 2.

Table 2. Mean and stand. dev. (sd) on simulated data of the percentage of original and spurious relations deleted by the repair process with $\alpha = 1$.

miRNAs	mRNAs	interactions	threshold	original (%)		spurious (%)	
				mean	sd	mean	sd
41	1,479	6,763	-0.3	5.6	3.2	8.7	4.5
			-0.35	22.5	4.9	34.8	7.6

In contrast to results on simulated noisy contexts, deletions affects both original and spurious relations. For both thresholds, the mean and standard deviation for spurious interactions is slightly higher than for the original interactions. A comparison between the two thresholds shows that the more stringent threshold has a higher mean and standard deviation. Interestingly, the same behaviour is observed when comparing simulated noisy contexts of size 40×40 with a noise probability of 0.01 or 0.05 (Table 1).

8 Conclusion

We have formalized the effect of noise on a microRNAs/mRNAs interaction graph by considering it has a formal context. Two types of noise may occur, namely spurious and missing relations. We showed that noise has the effect of increasing the set of original concepts linearly or exponentially respectively with respect to purely spurious or missing concepts. In most cases, there exists some intersection/inclusion relation between noisy concepts observable as a direct or a sibling relation in the modified concept lattice, which allows to recover the original concepts.

Based on these observations, two repair operations: *delete* and *add* have been defined for spurious and missing relations. These operations are applied on subsets of concept pairs, looking for the optimization of a score based on minimum description length principle. We have shown on simulated noisy

contexts that there exists a range of context sizes such that our method allows to increase the sensitivity of a highly specific prediction with the *delete* operation.

In order to test our method on more realistic data, we used a set of simulation parameters derived from a real miRNAs/mRNAs interaction graph. Unfortunately, the discriminative power of the repairing method on these data is insufficient as it deletes a significant number of true interactions.

Additional work is necessary to increase the deletion rate of spurious relations, to improve the discriminative power of the method for small and very large contexts and greater levels of noise. In the continuation of this work, we will also evaluate how well the *add* operation performs on the same data. Another perspective is to check how missing relations detected by our method compare to missing relations in approximated concepts, i.e. the 0 in dense rectangles of 1 (see section 3).

Acknowledgement

This work was founded by ANR project miRNAdapt and Région Bretagne. The authors thank R. Jullien, V. Picard and C. Galiez for constructive remarks on the paper.

References

BELOHLAVEK, R. and VYCHODIL, V. (2006): Replacing full rectangles by dense rectangles: concept lattices and attribute implications. In: *IEEE Information Reuse and Integration.* IEEE, 117-122.

BESSON, J., ROBARDET, C. and BOULICAUT, J. F. (2005). Mining formal concepts with a bounded number of exceptions from transactional data. In : *Knowledge Discovery in Inductive Databases.* Springer, Berlin-Heidelberg, 33-45.

BLACHON, S., PENSA, R. G., BESSON, J., ROBARTDET, C., BOULICAUT, J. F. and GANDRILLON, O. (2007): Clustering formal concepts to discover biologically relevant knowledge from gene expression data. *In silico biology, 7(4), 467-483.*

BRYAN, K., TERRILE, M., BRAY, I. M., DOMINGO-FERNANDÉZ, R., WATTERS, K. M., KOSTER, J., VERSTEEG, R. and STALLINGS, R. L. (2014): Discovery and visualization of miRNAmRNA functional modules within integrated data using bicluster analysis. *Nucleic Acids Research, 42(3), e17.*

CHIL, S. W., ZANG, J. B., MELE, A. and DARNELL R. B. (2009): Argonaute HITS-CLIP decodes microRNAmRNA interaction maps. *Nature, 460, 479-486.*

ENRIGHT, A. J., JOHN, B., GAUL, U., TUSCHL, T., SANDER, C. and Marks D. S. (2003): MicroRNA targets in Drosophila. *Genome biology, 5(1), R1-R1.*

GALLOT, A., SHIGENOBU, S., HASHIYAMA, T., JAUBERT-POSSAMAI, S. and TAGU, D. (2012): Sexual and asexual oogenesis require the expression of unique and shared sets of genes in the insect Acyrthosiphon pisum. *BMC genomics, 13(1), 76.*

GANTER, B., and WILLE, R. (1999): *Formal Concept Analysis: Mathematical Foundations.* Springer, Berlin.

GRIMSON, A., FARH, K. K. H., JOHNSTON, W. K., GARRETT-ENGELE, P., LIM, L. P., and BARTEL, D. P. (2007): MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Molecular cell, 27(1), 91-105.*

JANGA S. C. and VALLABHANENI S. (2011): MicroRNAs as Post-Transcriptional Machines and their Interplay with Cellular Networks. In: *RNA Infrastructure and Networks.* Springer, New-York, 59-74.

KLIMUSHKIN, M., OBIEDKOV, S. and ROTH, C. (2010): Approaches to the Selection of Relevant Concepts in the Case of Noisy Data. In: L. Kwuida and B. Sertkaya (Eds.): *Formal Concept Analysis.* Springer, Berlin-Heidelberg, 255-266.

REYES-HERRERA, P. H., FICARRA, E., ACQUAVIVA A. and MACII E. (2011): miREE: miRNA recognition elements ensemble. *BMC Bioinformatics, 12, 454-473.*

THE INTERNATIONAL APHID GENOMICS CONSORTIUM (2010): Genome Sequence of the Pea Aphid Acyrthosiphon pisum. *PLoS Biol, 8(2), e1000313.*

Physiological Entomology

Physiological Entomology (2013) **38**, 117–125 DOI: 10.1111/phen.12021

REVIEW ARTICLE

Genome expression control during the photoperiodic response of aphids

GAEL LE TRIONNAIRE, VALENTIN WUCHER and DENIS TAGU

INRA Rennes, UMR IGEPP, Le Rheu, France

Abstract. Aphids are major crop pests and show a high level of phenotypic plasticity. They display a seasonal, photoperiodically-controlled polyphenism during their life cycle. In spring and summer, they reproduce efficiently by parthenogenesis. At the end of summer, parthenogenetic individuals detect the transition from short nights to long nights, which initiates the production of males and oviparous females within their offspring. These are the morphs associated with the autumn season. Deciphering the physiological and molecular events associated with this switch in reproductive mode in response to photoperiodic conditions is thus of key interest for understanding and explaining the remarkable capacity of aphids to adapt to fluctuations in their environment. The present review aims to compile earlier physiological studies, focussing on the neuroendocrine control of seasonal photoperiodism, as well as a series of large-scale transcriptomic approaches made possible by the recent development of genomic resources for the model aphid species: the pea aphid *Acyrthosiphon pisum*. These analyses identify genetic programmes putatively involved in the control of the initial steps of detection and transduction of the photoperiodic signal, as well as in the regulation of the switch between asexual and sexual oogenesis within embryonic ovaries. The contribution of small RNAs pathways (and especially microRNAs) in the post-transcriptional control of gene expression, as well as the role of epigenetic mechanisms in the regulation of genome expression associated with the photoperiodic response, is also summarized.

Key words. Aphids, epigenetic mechanisms, neuroendocrine control, microRNAs, photoperiodism, transcriptomic analysis.

Introduction

Photoperiodism in the living world

Demographic success and the survival of most organisms is highly dependent on their ability to cope with and respond to a variety of environmental factors that can be either biotic or abiotic. Biotic factors mainly correspond to pathogens, parasites or predators that limit the development of a given species. Abiotic factors mainly include temperature, humidity and photoperiod, and their combined and continuous fluctuations across seasons can have a strong impact on the fitness of a wide range of organisms. To face the constant modifications of environmental factors, organisms have developed strategies enabling their long-term adaptation to season alternation. The most common and perhaps reliable mechanism is known as photoperiodism, where organisms can detect variations of day length that occur during the year and use them as signals to trigger the establishment of phenotypic/behavioural modifications, allowing their adaptation to seasons. Manifestations of photoperiodism are widespread amongst a variety of organisms, such as fungi, plants and animals. Plants synchronize their life cycle with season alternation. *Arabidopsis thaliana* is a facultative long-day plant because flowering is promoted by long days and delayed under short-day conditions. By contrast, flowering time in rice is induced by short days (Yanovsky & Kay, 2003). In numerous species of birds, reproduction and migration timing are controlled by endogenous circannual rhythmicity. The expression (or not) of such rhythms depend on the photoperiod and its fluctuation across seasons (Gwinner, 2003). Humans are also sensitive to photoperiod changes because afflictions such

Correspondence: Gaël Le Trionnaire, INRA Rennes, UMR 1349 IGEPP, BP 35327, 35657 Le Rheu, Cedex, France. Tel.: +33 2 2348 51 65; e-mail: gael.letrionnaire@rennes.inra.fr

as seasonal affective disorder can be diagnosed at the arrival of autumn and winter (Davis & Levitan, 2005). In invertebrates, insects are striking examples of organisms displaying a photoperiodic response. Cold tolerance, migration and growth rate regulation are common responses of insects to day-length changes. However, the two most striking examples of photoperiodism within the insects are diapause and the appearance of seasonal morphs. Diapause is an arrest of development during the life cycle of the insect that allows the anticipation of adverse environmental conditions (drought or cold) and is often triggered by photoperiod (Saunders *et al*., 2004). The production of seasonal morphs in aphids is historically documented (Tagu *et al*., 2008): the detection of short days induces a shift from clonal, viviparous reproduction (parthenogenesis) to sexual, oviparous reproduction. A switch from viviparous to oviparous embryogenesis thus occurs within the individuals that detect changes in day length (Tagu *et al*., 2005). At the population scale, lineages that are able to respond to photoperiodic cues coexist with lineages that have lost this ability and reproduce asexually during their life cycle (Simon *et al*., 2011). Depending on the type of organism, photoperiodic response can either result in a behavioural change or in the expression of a plasticity of the phenotype that will be more suited to the future environmental conditions. In this context, aphids represent an extreme case of phenotypic plasticity because the sexual morphs produced by asexual individuals that experience photoperiod changes correspond to different and contrasting phenotypes. This discrete phenotypic plasticity is also called polyphenism (Simpson *et al*., 2011). In aphids, photoperiodism is thus achieved by a plasticity of the reproductive mode. Understanding the molecular basis of this phenomenon offers the possibility not only to decipher the molecular mechanisms involved in the detection and transduction of the photoperiodic signal, but also to understand the molecular events governing the transition from an asexual to a sexual reproductive mode and embryogenesis. Aphids represent an ideal model for understanding the direct phenotypic consequences of the modification of photoperiod. The present review first introduces the aphid model and then focusses on the physiological and transcriptomic bases of key steps of this phenomenon such as the detection and transduction of the photoperiodic signal and the switch from asexual to sexual oogenesis within embryonic ovaries. The second part of this review focusses on the putative role of post-transcriptomic and epigenetic mechanisms associated with the establishment of phenotypic plasticity in response to the photoperiodic changes in aphids.

Aphids: major crop pests remarkably adapted to their environment

Aphids are phloem sap-feeding hemipterous insects that can cause significant economic losses on various crops such as wheat or maize. In temperate and continental regions, most aphid species reproduce quickly and efficiently by viviparous parthenogenesis during spring and summer. At the arrival of autumn, parthenogenetic individuals detect short days. Once sensed, this signal is transmitted to the embryos, which, in turn, direct their development towards becoming sexual adults. The sexual individuals produced (i.e. males and oviparous females) mate and females lay cold-resistant eggs that can withstand potentially adverse winter conditions (Fig. 1). The aphid genome is thus highly 'plastic' in the sense that it is able to predict and respond to environmental parameters (seasons) that can be strongly limiting for their survival and general fitness. Aphids are insects that are remarkably adapted to their environment by being able to respond to its fluctuation, explaining their success as one of the major crop pests. Understanding the molecular events regulating the photoperiodic response and more generally phenotypic plasticity in aphids is of major fundamental and agronomical interest for developing sustainable crop pest management strategies. The development of genomic resources within the aphid scientific community in the early years of the 21st Century has allowed significant progress. The pea aphid *Acyrthosiphon pisum* genome has recently been sequenced and annotated (Richards *et al*., 2010), which constitutes an absolute pre-requisite for a wide range of genetic and genomic analyses.

Physiological and transcriptomic bases of photoperiodism in aphids

Photoperiodic signal, photoperiodic clock and photoreceptors. Studies on photoperiodism in aphids started to emerge in the second part of the 20th Century, when it was demonstrated that aphids could measure scotophase (night length). For each species, there is a minimum length of scotophase (i.e. critical night length) above which the induction of sexual morphs is effective (Lees, 1973). A minimum number of consecutive long nights is also necessary to trigger the reproductive mode switch under controlled conditions. It can vary between species, although an average of at least ten consecutive long nights is sufficient to trigger the reproductive mode switch. This might be because day length is interpreted as an adaptive strategy to limit a too rapid switch that could be induced by a short period of exposure to long-nights. Complementary studies showed that temperature could also modulate the photoperiodic response (Lees, 1989). The nature of the photoperiodic clock involved in detecting short days has nevertheless not been clearly stated. The involvement of the circadian clock in the photoperiodic response in aphids has been questioned for a long-time. Three main theoretical models for the mechanism of insect photoperiodic clocks have been proposed, two of which suggest an involvement of the circadian clock (Internal or External Coincidence Model), whereas the third ('Hourglass' model) does not involve any circadian component. These models are at their essence theoretical and are still largely debated (Danks, 2005; Saunders, 2005). The accumulation of molecular evidence would thus clearly help in discriminating these different models. A recent study (Cortés *et al*., 2010) revealed the presence in the pea aphid genome of orthologues for several well-known *Drosophila* circadian clock genes such as *period*, *timeless*, *Clock*, *vrille* and *Pdp1*. Expression analyses confirmed a circadian rhythmicity for some of those

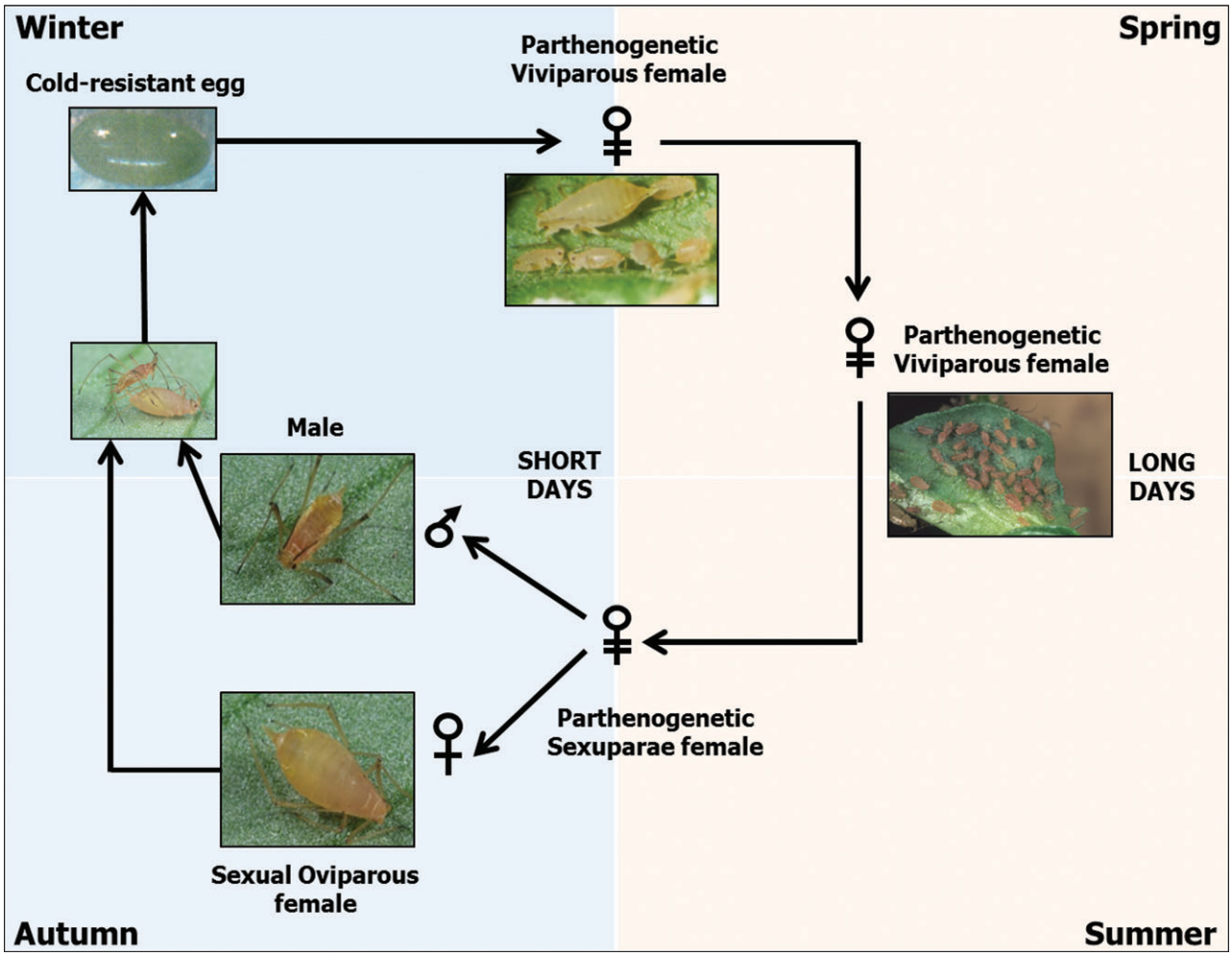


Fig. 1. Pea aphid life cycle and the production of seasonal morphs. In sexual lineages, aphids reproduce efficiently and quickly by parthenogenesis during spring and summer. At the end of summer, parthenogenetic individuals detect short days and initiate the production of sexual individuals in their offspring. Such individuals that produce sexual forms are also called 'sexuparae'. Autumn morphs (i.e. sexual females and males) are thus produced and will mate to produce cold-resistant eggs that can overcome winter.

genes, as well as a significant effect of photoperiod on the amplitude of oscillations. Nevertheless, the exact contribution of the circadian clock to the photoperiodic response remains unknown. The nature and localization of putative photoperiodic photoreceptors has also been investigated. Antibodies directed against a wide range of opsins and other phototransduction proteins were tested and shown to be localized in the ventral anterior region of the protocerebrum, suggesting that the photoperiodic photoreceptors could be located in this area of the brain (Gao *et al*., 1999). The molecular nature and the precise function of these receptors in the photoperiodic response remain unknown.

Neuroendocrine control. In insects, both endocrine glands and neurosecretory cells can release hormonal components. Steel & Lees (1977) showed that one of the five groups of neurosecretory cells from the protocerebrum (Cell Group I) was involved in the photoperiodic response because microcauterization of those cells abolished the response. These cells have long axons spreading into the abdomen of the aphid. Steel & Lees (1977) suggested that secretions (hormones or neuropeptides) from these cells are transported all along the axon and released at specific sites close to the ovarioles, which are the target tissues of the photoperiodic signal, although this has never been demonstrated. The nature of these neurosecretory molecules remains unknown. A recent combination of bioinformatics analyses, brain peptidomics and cDNA analyses allowed the establishment of a catalogue of pea aphid neuropeptides and neurohormones. Forty-two genes encoding neuropeptides and neurohormones were identified. The neuropeptides accumulated in the Group I of neurosecretory cells are probably rich in cysteine (because they respond to fuchsin staining). By correlating the type of neuropeptides rich in cysteine present in the pea aphid genome, and also the knowledge of neuropeptides secreted in other insects, it appears

that insulins could represent good candidates for neuropeptides involved in the regulation of photoperiodism (Huybrechts *et al*., 2010). This hypothesis appears to be realistic considering the results obtained from recent transcriptomic analyses of the photoperiodic response (Le Trionnaire *et al*., 2009), which show the differential expression of transcripts involved in the insulin signalling pathway (see below).

The involvement of Juvenile Hormones (JH) (known to regulate a wide range of developmental processes in insects) in the control of photoperiodism has also been studied. Topical application of JH or Kinoprene (a JH analogue) on the abdomen of viviparous aphids producing sexual individuals resulted in the reversion of the response to production of asexual individuals (Hardie & Lees, 1985). JH thus appears to play a role in the transduction of the photoperiodic signal. The role of melatonin in the photoperiodic response has also been investigated. In insects, this hormone is involved in the regulation of the visual system and displays a circadian rhythm of expression in head tissues (Bloch *et al*., 2012). Long-day, parthenogenetic aphids treated with this hormone produce sexual individuals in their offspring instead of asexual individuals (Gao & Hardie, 1997). These results indicate that melatonin might also play a role in the transduction of the photoperiodic signal. To further elucidate the molecular bases of the photoperiodic response, a fine analysis of the genetic programmes set up during this process was needed.

Genetic programmes associated with photoperiodic signal detection and transduction. Initial studies used methods such as the differential display reverse transcriptase-polymerase chain reaction (DD-RT-PCR) or suppression subtractive hybridization to identify transcripts differentially expressed between aphids reared under long days (producers of parthenogenetic progeny) and short days (producers of sexually-reproducing offspring). A transcript homologous to an amino acid transporter within GABAergic neurones was first identified by DD-RT-PCR as being over-expressed in short-day, sexual-offspring-producing individuals (Ramos *et al*., 2003). A putative role for this transcript in the transduction of the photoperiodic signal was proposed. Suppression subtractive hybridization approaches coupled with quantitative RT-PCR then allowed the identification of transcripts coding cuticular proteins and a β-tubulin that could play a role in hormone responses (Cortés *et al*., 2008). The precise function of these candidate genes in the regulation of photoperiodism is nevertheless unknown.

Genomic resources such as expressed sequence tag libraries from various aphid tissues were generated (Sabater-Muñoz *et al*., 2006). These libraries were used to build two generations of cDNA microarrays containing, respectively, 1700 (Le Trionnaire *et al*., 2007) and 7000 transcripts (Le Trionnaire *et al*., 2009, 2012). Heads of aphids reared under long-day or short-day photoperiods were collected at five stages of development during the process of sexual morph induction. By focusing on heads and cerebral tissues, the aim was to capture the genetic programmes set up during the initial steps of photoperiodic signal detection and transduction (Le Trionnaire *et al*., 2007, 2009). Microarray hybridizations combined with proteomics approaches (two dimensional differential in gel electrophoresis) revealed the differential expression of a significant number of transcripts (10% of spotted cDNAs) and peptides within the heads of aphids in response to short photoperiods, allowing the identification of several genetic programmes that could be associated with the photoperiodic response (Fig. 2). Among these, a subset of transcripts showed homologies with *Drosophila melanogaster* genes involved in the visual system such as *Arrestin* and *Calnexin*, known to play a role in rhodopsin phototransduction and maturation. This confirmed an earlier study showing that antibodies against a vertebrate arrestin strongly labelled the putative brain photoperiodic photoreceptors (Gao *et al*., 1999). Another set of transcripts were related to the nervous system, with several transcripts differentially expressed displaying homologies with *Drosophila* genes involved in axon guidance (*Rho I*, *NLaz*, *Capulet* and *Wunen*) and neurotransmission (*Kinesin*, *Dunc 10-4A*, *Dunc 13-4A* and a DEP-containing domain protein), strongly suggesting an involvement of the nervous system in the transduction of the photoperiodic signal. Insulin signalling might also play a role because one transcript encoding an insulin-degrading enzyme and another one coding for an insulin receptor were found to be differentially expressed in response to short photoperiods. Unexpectedly, a large number ($n = 38$) of cuticular protein transcripts appeared to be regulated. Most of them ($n = 25$) contained a RR domain (RR1 or RR2) that allows chitin-cuticular protein linkage (Gallot *et al*., 2010). Most of these transcripts were down-regulated under short-day photoperiods, suggesting a putative relaxing of the chitin-cuticular protein network in response to short days. Cuticle also contains N-β alanyl dopamine (NBAD) that allows linkage between cuticular proteins to produce hard-cuticle or sclerotization. NBAD is made of dopamine and β-alanine and the enzyme responsible for this conjugation is coded by the *ebony* gene. β-Alanine is synthesized from aspartate by the action of an enzyme coded by *black* gene. Transcriptomic analyses revealed that *ebony* and *black* transcripts were down-regulated in short-day reared aphids. Consequently, it can be hypothesized that less NBAD is synthesized under short-day conditions. This suggests that short photoperiods could result in the reduction of sclerotization level in the aphid heads, thereby modifying cuticle structure. These observations also raise the question of the level of dopamine in aphid heads under short-day conditions. Indeed, if less NBAD is synthesized, is the general level of dopamine affected? Dopamine synthesis involves two main enzymes: tyrosine hydroxylase (*th*), which metabolizes tyrosine into L-3,4-dihydroxyphenylalanine (L-DOPA), and dopa-decarboxylase (*ddc*), which metabolizes L-DOPA into dopamine. RT-PCR experiments showed that *th* and *ddc* transcripts were down-regulated in short-day reared aphid heads, suggesting that short photoperiods could result in a diminution of dopamine synthesis within aphid brains (Gallot *et al*., 2010). Because dopamine is a neurotransmitter (and a neurohormone), it is tempting to speculate that this molecule might be involved in the transduction of the photoperiodic signal. A recent study in *Locusta migratoria* demonstrated that the dopamine synthesis pathway was involved in the transition

from the solitary to the gregarious phase (Ma *et al*., 2011). More precisely, the data showed that *th* (tyrosine hydroxylase), *henna* and *vat1* (vesicle amino-acid transporter), three genes coding for enzymes involved in dopamine biosynthesis and synaptic release, were significantly down-regulated during the solitary phase. Functional and pharmacological analyses confirmed that the dopamine pathway was clearly involved in the behavioural transition (Ma *et al*., 2011). Because such a behavioural change in the locust is a case of phase polyphenism (but not triggered by day length changes), a clear parallel with reproductive polyphenism (triggered by photoperiod shortening) can be made and the dopamine biosynthesis pathway might also be involved in the transition from asexual to sexual reproduction in response to short days in aphids. To address this, the level of expression, the localization and the functional characterization of *pale*, *vat1* and *henna* transcripts in both long- and short-day reared aphids all have to be investigated. It is striking to emphasize that some of these transcriptomic modifications observed on aphids reared under controlled conditions were also detected in aphids reared outdoor under natural photoperiodic conditions. However, the differential expression of several heat-shock protein transcripts also suggested a strong response of aphids to additional environmental parameters such as temperature (Le Trionnaire *et al*., 2012).

Transcriptomic modifications associated with the transition from asexual to sexual oogenesis within embryonic ovaries. Once short days/long nights are detected by aphids, this signal has to be transduced to the target tissues, which are the embryos. A recent large-scale transcriptomic approach thus aimed to study the consequences of photoperiodic signal detection and transduction on embryo phenotypic plasticity. Transcriptomes from sexual and asexual embryos along a developmental series were compared using an oligo-nucleotide microarray with approximately 24 000 transcripts (Gallot *et al*., 2012). Based on previous studies (Corbitt & Hardie, 1985), a perfectly synchronized system was developed to target transcriptomic modifications associated only with the asexual to sexual oogenesis transition in the embryonic ovaries. Aphids reared under short photoperiods contain sexual embryos with a haploid meiotic germline. When Kinoprene (a JH analogue) is applied to the dorsal side of the abdomen, these embryos reverse their reproductive mode and produce asexual embryos containing a diploid non-meiotic germline. Under these conditions, sexual and asexual embryos are perfectly synchronized because the photoperiod does not change. This fine-tuned experimental design was used to compare the transcriptomes of asexual and sexual embryos at three stages of development: 18, 19 and 20 as defined by Miura *et al*. (2003). These are the final three developmental stages in aphid embryogenesis and correspond to eye differentiation (stage 18), muscle formation (stage 19) and the mature embryo (stage 20). Kinoprene treatment is performed when embryos are at stage 17 (i.e. the latest stage that responds to the hormonal treatment). After that specific stage, embryos are no longer responsive. This developmental window was chosen to study the direct effect of kinoprene on the sexual to asexual oogenesis switch. Statistical analysis of microarrays hybridization results revealed that only a few transcripts ($n = 33$) were differentially expressed between the two types of embryos. *In situ* hybridizations confirmed that most of the transcripts were located within germ cells and/or oocytes of asexual and/or sexual ovaries. Regulated transcripts could be assigned to four main functional categories (Fig. 2). Seven of those are involved in oogenesis, with a few playing a role in oocyte axis formation and specification (*orb* and *nudel*) or female meiosis chromosome segregation (*nanos*). Five transcripts play a role in post-transcriptional regulation, such as polyA-tail stabilization (*Pop2*). Four transcripts are also involved in epigenetic regulations (see below) and three in cell cycle control (*cyclin J*). These transcripts may therefore determine the aphid clonal or sexual oogenesis. It was thus revealed that JH signalling might control (directly or indirectly) the reproductive fate of aphid embryos.

Combined together, these large-scale transcriptomic approaches allowed the identification of a significant number of candidate transcripts that could play a key role in the detection and transduction of the photoperiodic signal, as well as in the transition from asexual to sexual oogenesis within embryonic ovaries (Fig. 2). The precise function of these different transcripts needs to be tested. The development of stable transgenesis tools remains challenging in aphids, mainly as a result of the complexity of the biological model (telescoping of generations, asexuality with lack of recombination events being predominant during the life cycle). So far, only transitory methods of transcripts silencing (RNA interference) have been developed in aphids with the direct injection of double-stranded RNAs into aphids (Mutti *et al*., 2006, 2008; Jaubert-Possamai *et al*., 2007; Shakesby *et al*., 2009) or by feeding aphids on plants expressing double-stranded RNAs in phloem sap (Pitino *et al*., 2011; Pitino & Hogenhout, 2013). These technologies displayed various levels of efficiency, mainly depending on the tissue localization of targeted transcripts. Pharmacological approaches (hormone or neurotransmitters injected or topically applied) appear to be a promising alternative for validating the function of specific candidate transcripts or at least signalling/biosynthetic pathways. Nevertheless, strong and efficient methods to modify gene expression in aphids are still missing.

Contribution of post-transcriptional and epigenetic mechanisms

The global expression of a genome is the result of a combination of transcriptomic and post-transcriptomic events that contribute to the establishment of a given phenotype. Small noncoding RNAs and especially microRNAs have emerged in the last years as key post-transcriptional regulators of gene expression (Kim *et al*., 2009). However, the expression of these different molecules (mRNAs and small RNAs) depends on the accessibility of corresponding genomic regions to transcriptional machinery or transcriptional modulators/regulators. This so-called ‘epigenetic’ state of the genome will thus be at

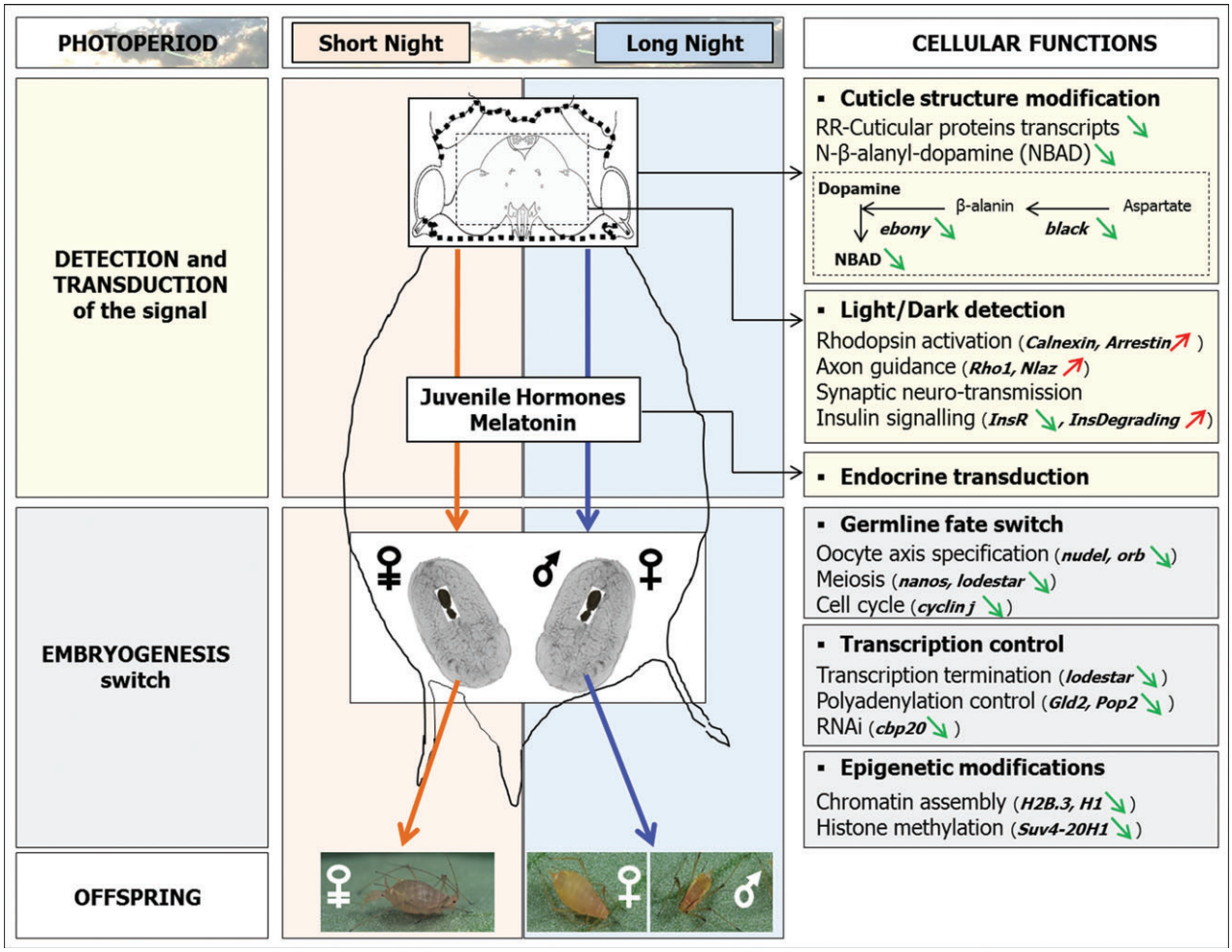


Fig. 2. Hypothetical model for regulation of seasonal photoperiodism in aphids. Recent large-scale transcriptomic analyses combined with earlier physiological studies allowed the identification of genetic programmes that might play key roles in the regulation of the photoperiodic response. The initial steps of detection and transduction of the photoperiodic signal appear to be associated with a modification of cuticle structure that could be linked to a reduction in dopamine levels within aphid heads. Visual and brain nervous systems might also play a role in this signalling step. Juvenile Hormones were also shown to play a central role in the endocrine transduction of this signal from the brain to the target tissues displaying the reproductive mode switch in the embryos. Later steps corresponding to a shift from asexual to sexual oogenesis appear to be associated with the differential expression of transcripts involved in germline fate and oogenesis, transcriptional and post-transcriptional control, as well as epigenetic modifications.

the basis of global genome expression and shape phenotypes. A given epigenome can be explained by a combination of DNA methylation patterns and chromatin structure and organization. Integrating post-transcriptional and epigenetic data with already well-identified transcriptomic changes associated with the photoperiodic response should thus allow the fine characterization of genome expression modifications associated with seasonal photoperiodism in aphids.

MicroRNAs and alternative morph production. The first catalogue of pea aphid microRNAs has been recently completed (Legeai *et al*., 2010). A combination of bioinformatic prediction of putative hairpin structures (typical of pre-microRNAs) on the genome and high-throughput sequencing of small RNAs from the whole bodies of parthenogenetic individuals allowed the identification of 149 microRNAs, including 55 conserved and 94 new microRNAs. The level of expression of candidate microRNAs between different aphid morphs (parthenogenetic females producing asexual progeny, oviparous/sexual females and parthenogenetic females producing sexual offspring, also called sexuparae) was then tested using a dedicated microRNA chip. Statistical analyses allowed the identification of 17 microRNAs (12 mature miRNAs and 5 miR*) displaying morph-specific profiles of expression. Seven microRNAs were differentially expressed between oviparous females and sexuparae, and nine were differentially expressed between oviparous and parthenogenetic females. Interestingly, ap-let-7 and ap-mir-100 were up-regulated in oviparous females

compared with parthenogenetic and sexuparae females. Their *Drosophila* homologues let-7 and miR-100 have been reported to play a role in metamorphosis and the response to ecdysone, a hormone involved in insect development. Ap-miR-34 also showed different expression levels between sexuparae and parthenogenetic females, which differ by the type of embryos they contain (sexual versus asexual). Interestingly, miR-34 is regulated in *D. melanogaster* by ecdysone as well as by JH. These microRNAs might thus target transcripts that could play key roles in morph specification and, by extension, in the photoperiodic response.

Sequencing and annotation of the pea aphid genome (IAGC, 2010) revealed that it displayed a high rate of gene duplication. For example, it shows an unexpected expansion of the microRNA pathway for genes that are highly conserved and have only a single copy in most organisms (Jaubert-Possamai *et al*., 2010; Ortiz-Rivas *et al*., 2012). There are indeed two copies of the microRNAs pathway-specific *dcr1* and *ago1* genes. One of the two copies (*dic1-b* and *ago1-b*) shows accelerated evolution. RT-PCR experiments also showed a morph-biased expression of these genes showing an accelerated evolution (e.g. *dic1-b* and *ago1-b*). This observation raises questions about the specific function of these duplicated copies in the microRNAs pathway within specific aphid morphs, especially in morphs displaying the reproductive mode switch. Further functional analysis will be needed to assess the specific roles of these duplicated copies.

However, systems biology could possibly leverage the lack of functional characterization. MicroRNAs and mRNAs work as a network of interactions because thousands of such interactions are usually predicted for one given species and one specific trait. Genes network and graphs methods are currently being developed to answer this question. A graph can integrate different information: microRNAs–mRNAs interactions, their differential level of expression between two conditions, and additional relationships, such as regulation by transcription factors. This integrated graph allows a global view of a given biological phenomenon. The constitution of such networks in the course of asexual to sexual oogenesis within embryonic ovaries might thus help identify new key regulators of photoperiodism in aphids.

DNA methylation in the pea aphid A. pisum

In mammals, DNA methylation is usually associated with promoter regions and highly methylated regions are correlated to low transcription. This methylation pattern is somehow different in insects. Even if some insect species such as beetles and *Drosophila* appear to have lost DNA methylation (Patalano *et al*., 2012), pea aphid as well as honey bee *Apis mellifera* or locust genome annotation confirmed that all the genes from the DNA methylation pathway are present (Walsh *et al*., 2010; Hunt *et al*., 2010). Methylation appears to be important in social insects such as honey bees, which also exhibit a phenotypic plasticity (caste morphs). A recent study showed that 550 genes displayed a differential methylation pattern between queens and workers. Strong correlations between methylation patterns and splicing sites were also found. It was proposed that modulation of alternative splicing could be one of the mechanisms by which DNA methylation is linked to gene regulation in the context of phenotypic plasticity (Lyko *et al*., 2010). In the case of the pea aphid, Walsh *et al*. (2010) showed that 0.69% of all cytosines were methylated. Methylation appears to be restricted to gene coding sequences at CpG sites. The precise role of DNA methylation in reproductive mode plasticity in response to photoperiod has not been studied yet, although some studies are currently underway aiming to analyse the role of DNA methylation in the regulation of dispersal polyphenism (Srinivasan & Brisson, 2012). It would thus be of great interest to evaluate the contribution of this epigenetic pathway to the regulation of photoperiodism by evaluating in details DNA methylation patterns between morphs.

Chromatin organization and histone modifications. Chromatin is defined as the association between DNA and proteins (histones and nonhistone proteins). Nucleosomes are sub-units of chromatin made of a DNA fragment of 140 bp wrapped around a protein complex of two copies of each histone protein (H2A, H2B, H3 and H4). Nucleosome numbers and organization all along the chromosome can shape accessibility of genomic regions such as promoters to transcription factors or other regions such as enhancers to regulatory molecules. Nucleosome occupancy can be studied by recently developed methods such as FAIRE-seq (formaldehyde-associated isolation of regulatory elements; Kaplan *et al*., 2008) and MAINE-seq (MNase-mediated purification of mononucleosomes; Simon *et al*., 2012) that allow the isolation of protein-free DNA and histone-bound DNA, respectively. Such methods are of great interest for identifying genomic regions epigenetically regulated during a given phenomenon. Nucleosomic histones can also be modified post-translationally. Histone residues such as specific lysines (K) can be methylated or acetylated. The combination of different histone modifications will have consequences for the level of DNA accessibility. Different chromatin states can then be defined by a combination of several histone modification marks. For example, genome-wide profiling of a combinatorial pattern of enrichment or depletion for specific histone modification marks has been established for all the chromosomes of *Drosophila*, allowing the establishment of a nine-state model for *Drosophila* chromatin (Kharchenko *et al*., 2011). So far in aphids, only H3K9me mark and HP1 proteins have been localized on heterochromatic regions (Mandrioli & Borsatti, 2007). More recently, it has been shown that the pea aphid genome possesses a complement of metazoan histone-modifying enzymes with greater gene family diversity than that seen in a number of other arthropods. Several genes have undergone recent duplication and divergence, potentially enabling greater combinatorial diversity among the chromatin-remodelling complexes (Rider *et al*., 2010). The comparison of sexual and asexual aphid transcriptomes (Gallot *et al*., 2012) demonstrated the differential expression of transcripts coding proteins involved in epigenetic mechanisms, such as Histones H2B.3 and H1, which are known to participate to chromatin

assembly and disassembly. Another example concerns *Suv4-20H1*, which is involved in histone methylation. This fine comparison of sexual and asexual embryos transcriptomes already suggests that some epigenetic regulations involving chromatin structure modifications are occurring during phenotypic plasticity. Depicting the type of histone modifications associated with the reproductive mode switch of embryos in response to photoperiodic cues would thus be of great interest.

Perspectives

The regulation of photoperiodism in aphids and its effects on the embryo phenotypic plasticity has been extensively studied at the transcriptomic level. These large-scale studies have allowed the identification of some of the genetic programmes involved in the photoperiodic signal detection and transduction and in the embryos' reproductive mode switch. These studies have established an extensive catalogue of transcripts, hormones and neurotransmitters (e.g. insulin, dopamine) as candidates for further functional and pharmacological validation experiments. The recent and on-going development of high-throughput sequencing technologies now allows the identification of key post-transcriptional regulators of gene expression (such as microRNAs), as well as the mapping of distinct epigenetic marks (nucleosome occupancy, histone modification marks and DNA methylation patterns). The establishment of alternative phenotypes in response to environmental cues such as photoperiod definitely involves a combination of epigenetic, transcriptomic and post-transcriptomic regulatory events. An integrative view [in accordance with the modENCODE model (Celniker *et al*., 2009) but for a non-model organism such as aphids] of the contribution of these different mechanisms thus appears to be an ideal approach that should allow the identification of key genomic regions involved in the regulation of phenotypic plasticity, especially in the case of the aphid photoperiodic response.

Acknowledgements

Jennifer Brisson (University of Nebraska) is sincerely thanked for her help in reading and correcting this manuscript.

References

Bloch, G., Hazan, E. & Rafaeli, A. (2012) Circadian rhythms and endocrine functions in adult insects. *Journal of Insect Physiology*, **59**, 56–69.

Celniker, S.E., Dillon, L.A.L., Gerstein, M.B. *et al*. (2009) Unlocking the secrets of the genome. *Nature*, **459**, 927–930.

Corbitt, T.S. & Hardie, J. (1985) Juvenile hormone effects on polymorphism in the pea aphid, *Acyrtosiphon pisum*. *Entomological Experimental Application*, **38**, 131–135.

Cortés, T., Tagu, D., Simon, J. *et al*. (2008) Sex versus parthenogenesis: a transcriptomic approach of photoperiod response in the model aphid *Acyrthosiphon pisum* (Hemiptera: Aphididae). *Gene*, **408**, 146–156.

Cortés, T., Ortiz-Rivas, B. & Martínez-Torres, D. (2010) Identification and characterization of circadian clock genes in the pea aphid *Acyrthosiphon pisum*. *Insect Molecular Biology*, **19**, 123–139.

Danks, H. (2005) How similar are daily and seasonal biological clocks? *Journal of Insect Physiology*, **51**, 609–619.

Davis, C. & Levitan, R.D. (2005) Seasonality and seasonal affective disorder (SAD): an evolutionary viewpoint tied to energy conservation and reproductive cycles. *Journal of Affective Disorders*, **87**, 3–10.

Gallot, A., Rispe, C., Leterme, N. *et al*. (2010) Cuticular proteins and seasonal photoperiodism in aphids. *Insect Biochemistry and Molecular Biology*, **40**, 235–240.

Gallot, A., Shigenobu, S., Hashiyama, T. *et al*. (2012) Sexual and asexual oogenesis require the expression of unique and shared sets of genes in the insect *Acyrthosiphon pisum*. *BMC Genomics*, **13**, 76.

Gao, N. & Hardie, J. (1997) Melatonin and the pea aphid, *Acyrthosiphon pisum*. *Journal of Insect Physiology*, **43**, 615–620.

Gao, N., von Schantz, M., Foster, R.G. & Hardie, J. (1999) The putative brain photoperiodic photoreceptors in the vetch aphid, *Megoura viciae*. *Journal of Insect Physiology*, **45**, 1011–1019.

Gwinner, E. (2003) Circannual rhythms in birds. *Current Opinion in Neurobiology*, **13**, 770–778.

Hardie, J. & Lees, A. (1985) The induction of normal and teratoid viviparae by a juvenile hormone and kinoprene in two species of aphids. *Physiological Entomology*, **10**, 65–74.

Hunt, B.G., Brisson, J.A., Soojin, V.Y. & Goodisman, M.A. (2010) Functional conservation of DNA methylation in the pea aphid and the honeybee. *Genome biology and evolution*, **2**, 719.

Huybrechts, J., Bonhomme, J., Minoli, S. *et al*. (2010) Neuropeptide and neurohormone precursors in the pea aphid, *Acyrthosiphon pisum*. *Insect Molecular Biology*, **19**, 87–95.

Jaubert-Possamai, S., Le Trionnaire, G., Bonhomme, J. *et al*. (2007) Gene knockdown by RNAi in the pea aphid *Acyrthosiphon pisum*. *BMC Biotechnology*, **7**, 63.

Jaubert-Possamai, S., Rispe, C., Tanguy, S. *et al*. (2010) Expansion of the miRNA pathway in the hemipteran insect *Acyrthosiphon pisum*. *Molecular Biology and Evolution*, **27**, 979–987.

Kaplan, N., Moore, I.K., Fondufe-Mittendorf, Y. *et al*. (2008) The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature*, **458**, 362–366.

Kharchenko, P.V., Alekseyenko, A.A., Schwartz, Y.B. *et al*. (2011) Comprehensive analysis of the chromatin landscape in *Drosophila melanogaster*. *Nature*, **471**, 480–485.

Kim, V.N., Han, J. & Siomi, M.C. (2009) Biogenesis of small RNAs in animals. *Nature Reviews Molecular Cell Biology*, **10**, 126–139.

Le Trionnaire, G., Jaubert, S., Sabater-Munoz, B. *et al*. (2007) Seasonal photoperiodism regulates the expression of cuticular and signalling protein genes in the pea aphid. *Insect Biochemistry and Molecular Biology*, **37**, 1094–1102.

Le Trionnaire, G., Francis, F., Jaubert-Possamai, S. *et al*. (2009) Transcriptomic and proteomic analyses of seasonal photoperiodism in the pea aphid. *BMC Genomics*, **10**, 46.

Le Trionnaire, G., Jaubert-Possamai, S., Bonhomme, J. *et al*. (2012) Transcriptomic profiling of the reproductive mode switch in the pea aphid in response to natural autumnal photoperiod. *Journal of Insect Physiology*, **12**, 1517–1524.

Lees, A.D. (1973) Photoperiodic time measurement in the aphid *Megoura viciae*. *Journal of Insect Physiology*, **19**, 2279–2316.

Lees, A.D. (1989) The photoperiodic responses and phenology of an english strain of the pea aphid, *Acyrthosiphon pisum*. *Ecological Entomology*, **14**, 69–78.

Legeai, F., Rizk, G., Walsh, T. *et al*. (2010) Bioinformatic prediction, deep sequencing of microRNAs and expression analysis during

phenotypic plasticity in the pea aphid, *Acyrthosiphon pisum*. *BMC Genomics*, **11**, 281.

Lyko, F., Foret, S., Kucharski, R. *et al*. (2010) The honey bee epigenomes: differential methylation of brain DNA in queens and workers. *PLoS Biology*, **8**, e1000506.

Ma, Z., Guo, W., Guo, X. *et al*. (2011) Modulation of behavioral phase changes of the migratory locust by the catecholamine metabolic pathway. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 3882–3887.

Mandrioli, M. & Borsatti, F. (2007) Analysis of heterochromatic epigenetic markers in the holocentric chromosomes of the aphid Acyrthosiphon pisum. *Chromosome research*, **15**, 1015–22.

Miura, T., Braendle, C., Shingleton, A. *et al.* (2003) A comparison of parthenogenetic and sexual embryogenesis of the pea aphid Acyrthosiphon pisum (Hemiptera : Aphidoidea). *Journal of Experimental Zoology Part B-Molecular and Developmental Evolution*, **295B**, 59–81.

Mutti, N.S., Park, Y., Reese, J.C. & Reeck, G.R. (2006) RNAi knockdown of a salivary transcript leading to lethality in the pea aphid, *Acyrthosiphon pisum*. *Journal of Insect Science*, **6**, 38.

Mutti, N.S., Louis, J., Pappan, L.K. *et al*. (2008) A protein from the salivary glands of the pea aphid, *Acyrthosiphon pisum*, is essential in feeding on a host plant. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 9965–9969.

Ortiz-Rivas, B., Jaubert-Possamai, S., Tanguy, S. *et al*. (2012) Evolutionary study of duplications of the miRNA machinery in aphids associated with striking rate acceleration and changes in expression profiles. *BMC Evolutionary Biology*, **12**, 216.

Patalano, S., Hore, T.A., Reik, W. & Sumner, S. (2012) Shifting behaviour: epigenetic reprogramming in eusocial insects. *Current Opinion in Cell Biology*, **24**, 367–373.

Pitino, M. & Hogenhout, S.A. (2013) Aphid protein effectors promote aphid colonization in a plant species-specific manner. *Molecular Plant–Microbe Interactions*, **26**, 130–139.

Pitino, M., Coleman, A.D., Maffei, M.E. *et al*. (2011) Silencing of aphid genes by dsRNA feeding from plants. *PLOS ONE*, **6**, e25709.

Ramos, S., Moya, A. & Martínez-Torres, D. (2003) Identification of a gene overexpressed in aphids reared under short photoperiod. *Insect Biochemistry and Molecular Biology*, **33**, 289–298.

Richards, S., Gibbs, R.A., Gerardo, N.M. *et al*. (2010) Genome sequence of the pea aphid *Acyrthosiphon pisum*. *PLoS Biology*, **8**, e1000313.

Rider, S. Jr., Srinivasan, D. & Hilgarth, R. (2010) Chromatin-remodelling proteins of the pea aphid, *Acyrthosiphon pisum* (Harris). *Insect Molecular Biology*, **19**, 201–214.

Sabater-Muñoz, B., Legeai, F., Rispe, C. *et al*. (2006) Large-scale gene discovery in the pea aphid *Acyrthosiphon pisum* (Hemiptera). *Genome Biology*, **7**, 21.

Saunders, D.S. (2005) Erwin Bünning and Tony Lees, two giants of chronobiology, and the problem of time measurement in insect photoperiodism. *Journal of Insect Physiology*, **51**, 599–608.

Saunders, D.S., Lewis, R.D. & Warman, G.R. (2004) Photoperiodic induction of diapause: opening the black box. *Physiological Entomology*, **29**, 1–15.

Shakesby, A., Wallace, I., Isaacs, H. *et al.* (2009) A water-specific aquaporin involved in aphid osmoregulation. *Insect Biochemistry and Molecular Biology*, **39**, 1–10.

Simon, J.C., Pfrender, M.E., Tollrian, R. *et al*. (2011) Genomics of environmentally induced phenotypes in 2 extremely plastic arthropods. *Journal of Heredity*, **102**, 512–525.

Simon, J.M., Giresi, P.G., Davis, I.J. & Lieb, J.D. (2012) Using formaldehyde-assisted isolation of regulatory elements (FAIRE) to isolate active regulatory DNA. *Nature Protocols*, **7**, 256–267.

Simpson, S.J., Sword, G.A. & Lo, N. (2011) Polyphenism in insects. *Current Biology*, **22**, 352–363.

Srinivasan, D.G. & Brisson, J.A. (2012) Aphids: a model for polyphenism and epigenetics. *Genetics Research International*, **2012**, 431531.

Steel, C.G.H. & Lees, A.D. (1977) The role of neurosecretion in the photoperiodic control of polymorphism in the aphid *Megoura viciae*. *Journal of Experimental Biology*, **67**, 117–135.

Tagu, D., Sabater-Munoz, B. & Simon, J.C. (2005) Deciphering reproductive polyphenism in aphids. *Invertebrate Reproduction & Development*, **48**, 71–80.

Tagu, D., Klingler, J.P., Moya, A. & Simon, J.-C. (2008) Early progress in aphid genomics and consequences for plant-aphid interactions studies. *Molecular Plant-Microbe Interactions*, **21**, 701–708.

Walsh, T.K., Brisson, J.A., Robertson, H.M. *et al*. (2010) A functional DNA methylation system in the pea aphid, *Acyrthosiphon pisum*. *Insect Molecular Biology*, **19**, 215–228.

Yanovsky, M.J. & Kay, S.A. (2003) Living by the calendar: how plants know when to flower. *Nature Reviews Molecular Cell Biology*, **4**, 265–275.

Accepted 22 March 2013

Résumé

Cette thèse cherche à discriminer au niveau génomique entre le développement d'embryons vers un mode de reproduction sexué et le développement vers un mode asexué chez le puceron du pois, *Acyrthosiphon pisum*. Cette discrimination passe par la création du réseau de régulation post-transcriptionnelle des microARN et des ARNm qui possèdent des cinétiques d'expression différentes entre ces deux embryogenèses ainsi que par l'analyse des modules d'interactions de ce réseau par l'utilisation de l'analyse de concepts formels. Pour ce faire, une stratégie en plusieurs étapes a été mise en place : la création d'un réseau d'interactions entre les microARN et les ARNm du puceron du pois ; l'extraction et la réduction du réseau aux microARN et ARNm qui possèdent des cinétiques différentes entre les deux embryogenèses à partir des données d'expression tirées du séquençage haut-débit ; l'analyse du réseau d'interactions réduit aux éléments d'intérêt par l'analyse de concepts formels. L'analyse du réseau a permis l'identification de différentes fonctions potentiellement importantes comme l'ovogenèse, la régulation transcriptionnelle ou encore le système neuroendocrinien. En plus de l'analyse du réseau, l'analyse de concepts formels a été utilisée pour définir une méthode de réparation de graphe biparti basée sur une topologie en « concepts » ainsi qu'une méthode de visualisation de graphes bipartis par ses concepts.

Abstract

This thesis aims to discriminate between embryos development towards either sexual or asexual reproduction types in pea aphids, *Acyrthosiphon pisum*, at the genomic level. This discrimination involves the creation of a post-transcriptional regulation network between microRNAs and mRNAs whose kinetic expressions change depending on the embryogenesis. It also involves a study of this network's interaction modules using formal concept analysis. To do so, a three-step strategy was set up. First the creation of an interaction network between the pea aphid's microRNAs and mRNAs. The network is then reduced by keeping only microRNAs and mRNAs which possess differential kinetics between the two embryogeneses, these are obtained using high-throughput sequencing data. Finally the remaining network is analysed using formal concept analysis. Analysing the network allowed for the identification of several functions of potential interest such as oogenesis, transcriptional regulation or even neuroendocrine system. In addition to network analysis, formal concept analysis was used to create a new method to repair a bipartite graph based on its topology and a method to visualise a bipartite graph using its formal concepts.