COLLEGE ECOLOGIE
DOCTORAL GEOSCIENCES
BRETAGNE AGRONOMIE ALIMENTATION

L'INSTITUT agro Rennes Angers

THÈSE DE DOCTORAT DE

L'INSTITUT AGRO RENNES ANGERS

ÉCOLE DOCTORALE N° 600
Écologie, Géosciences, Agronomie, Alimentation
Spécialité : *Génétique, génomique et bio-informatique*

Par
Camille Juigné

Integration and analysis of heterogeneous biological data through multilayer graph exploitation to gain deeper insights into feed efficiency variations in growing pigs

Thèse présentée et soutenue à Rennes, le 1er décembre 2023
Unité de recherche : Physiologie, Envionnement et Génétique pour l'Animal et les Systèmes d'Élevage (UMR Pegase)

Rapportrice et rapporteur avant soutenance :

Andrea RAU	Directrice de recherche, INRAE
Fabien JOURDAN	Directeur de recherche, INRAE

Composition du Jury :

Président :	Mathieu EMILY	Professeur, Institut Agro Rennes-Angers
Examinateur :	Michel DUMONTIER	Distinguished Professor, Maastricht University
Rapportrice et rapporteur :	Andrea RAU	Directrice de recherche, INRAE
	Fabien JOURDAN	Directeur de recherche, INRAE
Dir. de thèse :	Florence GONDRET	Directrice de recherche, INRAE
Co-dir. de thèse :	Emmanuelle BECKER	Maîtresse de conférence, HDR, Université de Rennes

ACKNOWLEDGMENT, REMERCIEMENTS

Ecrire les remerciements, c'est un peu comme rédiger l'état de l'art : il est bien difficile de trouver le point de départ idéal, d'organiser les idées de manière cohérente et surtout, de ne rien oublier d'essentiel. Cependant, cette fois, il s'agit non pas, de dépeindre le paysage technique qui a encadré ces travaux mais de faire honneur au paysage relationnel qui a mené à son achèvement.

En premier lieu je tiens à remercier chaleureusement Andrea Rau, Fabien Jourdan, Michel Dumontier et Mathieu Emily pour l'intérêt qu'elle et ils ont porté à mon travail à travers des discussions enrichissantes et en acceptant de faire partie de mon jury de thèse. Je remercie tout particulièrement Andrea et Fabien pour le temps accordé à réviser ce manuscrit et pour leurs retours bienveillants et instructifs.

Je remercie inconditionnellement mes encadrantes, Florence Gondret et Emmanuelle Becker, pour cette aventure de trois ans que nous avons partagée. À Florence, je suis reconnaissante pour sa disponibilité constante au fil de ces trois années et son engagement dans ce projet, qui s'est révélé, sans doute, plus axé sur l'informatique qu'elle ne l'avait imaginé. À Emmanuelle, merci de m'avoir guidée dans le monde de la recherche et de la bio-informatique depuis 2019.
Avec vous, j'ai pu prendre des décisions de recherche en toute liberté, suivre mes intuitions et mes aspirations, tout en bénéficiant de votre soutien et de vos conseils quand j'en ai eu besoin. Merci pour ce précieux partage d'expérience.

Cette thèse n'aurait été la même sans la collaboration, la présence et le soutien sans faille d'Olivier Dameron. Tout ce que je pourrais écrire ne saurait retranscrire ma reconnaissance. Merci de m'avoir, souvent, donné et redonné confiance en moi et en nos travaux.

Un sincère remerciement à Yuna Blum, Christine Brun et Mathieu Emily pour avoir fait partie de mon comité de suivi de thèse, pour nos discussions et leurs conseils éclairés.

Certaines enseignantes et certains enseignants ont eu un impact sur moi-même et la trajectoire me menant jusqu'à cette thèse. A commencer par Yannick Le Bras, Pierre Hunger, Sylvain Tacquet, Sandrine Coicadan, Ludovic Menguy et Jean-Marie Poublanc, mes enseignants et enseignante de Classe Préparatoire aux Grandes Ecoles qui m'ont donné le goût insatiable de la rigueur scientifique.

Attirée par l'informatique et la côte bretonne, j'ai de nouveau croisé le chemin d'enseignantes et enseignants dont les qualités humaines et les enseignements inspirants m'ont poussée à approfondir leurs disciplines. En particulier, Damien Lolive en informatique formelle, Gwénolé Lecorvé en apprentissage automatique, qui, avec Claire Lepage, a joué un rôle dans l'élaboration de mon projet professionnel en bio-informatique en soutenant mon projet d'échange académique à l'Université du Québec à Montréal, ainsi qu'en plaçant l'équipe Dyliss sur mon chemin, François Goasdoué en algorithmique et base de données, que j'ai eu grand plaisir à croiser de nouveau lors de cette thèse.
Le tournant décisif entre mes études d'ingénieure et mon parcours académique a pris forme lors d'un stage au sein de l'équipe Dyliss, aux côtés d'Emmanuelle Becker, Gwenaël Rabut et Olivier Dameron. Quelle aubaine d'avoir croisé votre chemin à tous les trois ! J'ai appris grâce à la confiance que vous m'avez accordée à m'épanouir dans la recherche. Avec regrets, je quittais les couloirs de Symbiose, mais pour une courte durée grâce à François Moreews. C'est lui qui, en tant que collègue avec qui j'ai eu le plaisir de travailler en tant qu'ingénieure, m'a suggéré de répondre à cette offre de thèse, seulement 2 ou 3 jours avant la date limite de candidature. Thèse qui, quelques mois plus tard, deviendra *mon* sujet de thèse.

Un immense merci à toutes et tous les membres de Symbiose avec qui j'ai passé quatre très belles années. Une pensée spéciale à mon ami Kévin Da Silva dont la rencontre a été un véritable privilège et avec qui j'ai eu le plaisir de partager cette expérience. Je tiens également à exprimer ma gratitude envers Anthony Bretaudeau et son humour que j'ai adorés côtoyer au quotidien. De manière général, un énorme merci à chacune et chacun, pour avoir contribué à créer un environnement agréable qui me manquera. Que ce soit la présence rayonnante de Marie, les blagues d'Anthony, les conseils d'Emeline et de Catherine, les sourires de Jeanne, les repas et discussions partagés à Supélec, la gentillesse de Jacques, la bonne humeur de Pierre et de Karel, les défis et les conversations sportives de Stéphanie et Yann, les repas à la cafétéria autour de bocaux, les discussions et sorties entre doctorantes, doctorants et jeunes ingénieures et ingénieurs, le bureau D152 partagé avec Gildas et Matthieu, jusqu'à notre passion commune des pauses sublimées par des dégustations de mets divers et variés, allant de bons gâteaux à la douteuse marmite anglaise.

Cette thèse en collaboration entre l'INRAE et l'INRIA, m'a donné l'occasion de faire de nombreuses rencontres enrichissantes, notamment au sein de l'UMR Pegase. Je tiens à exprimer ma gratitude envers toutes les personnes qui ont manifesté de la curiosité et de l'intérêt pour mon sujet de recherche, qui se démarque par son caractère plus informatique que les thématiques

habituelles. Un merci spécial aux plus jeunes de l'unité, Elise, Angélique, Emma, Maeva, Jean-Charles, Lorry, Clément, Chloé, Ellyn, Alyson et à toutes les amatrices et amateurs du café au 84. Merci pour tous les bons moments partagés, les sorties, les courses, les trails, les festivals. Heureusement nous n'avons pas eu à bien comprendre nos sujets de thèse respectifs pour bien se comprendre sur ces passions communes.

Je suis reconnaissante d'avoir eu l'opportunité de faire de belles rencontres, que ce soit lors de l'école d'été MDD ou dans mes divers engagements en faveur de l'égalité, la diversité, et l'inclusion. J'ai notamment eu le plaisir de rencontrer des personnes courageuses et inspirantes qui œuvrent pour faire bouger les choses au sein de la commission égalité de l'Irisa. Egalement en participant à l'initiative 'Elles Codent Elles Créent' pour initier des collégiennes à Python, ainsi qu'aux échanges entre jeunes scientifiques sur les inégalités en sciences au sein du cercle de lecture que j'ai eu le plaisir de coorganiser avec mon amie Véronne Yepmo, ces moments ont été des occasions précieuses qui ont renforcé ma détermination à contribuer à un monde scientifique plus équitable et inclusif.

Un grand merci également à mes étudiantes et étudiants avec qui j'ai découvert les joies de l'enseignement et du partage des connaissances.

Finalement, cette expérience n'aurait pas été la même sans les à-côtés de la thèse. J'ai la chance d'avoir été très bien entourée : merci à mes amies et amis de toujours, Marion Simon, Juliette Brown, Justine Hermenier, Lou Richard, Maëlle Guillois, Philippine Dodin, Thibaut Charpentier, Léo Lechevalier, Axel Michaud ; à mes amies et ami de Bellevue, Enora Perrenou, Océana Renoult, Anne-Lise Duroy, Amélie Fretault et Antoine Raffray ; à mes chères amies et amis de Bretagne, Bachir El Atlas, Gwendal Thomas, Fannie Tamalet, Domitille Schanne ; à mes fidèles compagnons de concerts et festivals de Montesquieu, Paul Chable, Paul Tremoureux et Florent Papini ; et à mes partenaires sportives et sportifs : de la danse, du taekwondo, de l'escalade, des courses à pied et de trail, du renforcement musculaire de Thierry, des randonnées et des beaux treks.

Pour finir, un immense merci à ma famille : maman, papa, mamie, Pauline, Lucas, Sylvie, Nico, Anaïs, Romane, Raphaël, Paul et bien sûr Luna. Merci pour votre soutien inconditionnel.

En conclusion, je tiens à exprimer ma profonde gratitude envers toutes les personnes qui ont contribué à la réussite de cette thèse. Leurs soutiens précieux, conseils, encouragements et collaborations ont été la clé de cette aventure académique. À chacun d'entre vous, mes plus sincères remerciements.

RÉSUMÉ EN FRANÇAIS

Contexte

Les récentes avancées technologiques en acquisition de données biologiques entraînent une croissance exponentielle de données multimodales (de types différents) générées dans divers laboratoires, expériences et conditions. Ce volume de données en expansion rapide pose des défis scientifiques en termes de stockage, de standardisation et d'analyse [1, 2].

Cette thèse aborde spécifiquement les deux derniers défis, visant à maximiser l'utilisation des données issues d'expériences et à exploiter les connaissances disponibles dans diverses bases de données biologiques et ontologies. L'objectif est d'obtenir de nouvelles perspectives à partir de l'analyse approfondie des données, en tenant compte du contexte des connaissances stockées dans les bases de données et ontologies, afin d'améliorer notre représentation et compréhension de la biologie des systèmes.

Nous avons émis l'hypothèse qu'adopter une perspective holistique sur l'organisation biologique peut conduire à une meilleure compréhension des processus complexes en représentant les nombreuses relations entre les différentes entités biologiques.

Cela implique de tirer parti de diverses modalités -ome, telles que le génome, le transcriptome, le protéome, le métabolome, etc., en établissant des liens entre eux et en couvrant certains biais (sélectivité, sensibilité, etc.) associés aux technologies d'acquisition de données pour ces différents types d'entités. Le domaine des -omiques (c'est-à-dire les technologies pour acquérir des données liées aux différents niveaux -ome) englobe une large gamme d'objectifs, notamment la compréhension, la prévention, le diagnostic et la gestion des maladies, ainsi que l'optimisation des performances.

La plupart du temps, les changements mesurables dans des mélanges biologiques hautement complexes (cellules, tissus, fluides) à travers des conditions expérimentales ou environnementales sont analysés à un niveau d'organisation biologique donné (unimodal) à l'aide de méthodes statistiques univariées ou multivariées, et sont comparés avec les connaissances existantes grâce à une expertise spécialisée. Des méthodes pour l'analyse conjointe de données -omiques multimodales ont également été développées pour faciliter la cartographie des voies modifiées. Ces méthodes se concentrent principalement sur la réduction de la dimensionnalité des données en identifiant les caractéristiques communes importantes et en facilitant la manipu-

lation des données. Cependant, ces approches ne tiennent souvent pas compte des connaissances disponibles concernant les relations physiques entre les molécules (activation, inhibition, formation de complexes, interaction, etc.). Cet oubli peut entraîner la perte d'informations précieuses et l'absence d'informations ou d'interdépendances essentielles. Pour relever ces défis, nous proposons une approche d'intégration globale et systémique pour l'analyse des données -omiques multimodales qui prend en compte les connaissances présentes dans les bases de données et les bases de connaissances.

Il existe une multitude de bases de données spécialisées, chacune se concentrant sur des types spécifiques de données, utilisant différents formats et ayant ses propres processus de curation. En conséquence, ces bases de données sont complémentaires et leur intégration reste un défi important en raison de leur hétérogénéité et de l'existence de recouvrements potentiels entre elles. Notamment, il existe plusieurs bases de données dédiées à la description des voies métaboliques, qui facilitent l'établissement de liens entre les entités biologiques à travers les différentes couches -omiques. Pour relever ce défi d'intégration, la format Biological Pathway Exchange (BioPAX) s'est imposée comme une solution judicieuse. Ce format offre une représentation d'une grande finesse des connaissances sur les processus biologiques aux niveaux moléculaire et cellulaire. BioPAX permet l'intégration de diverses bases de connaissances liées aux voies métaboliques et facilite les références croisées avec des ressources externes, notamment des ontologies bien établies telles que UniProt pour les protéines, ChEBI pour les petites molécules, MI pour l'annotation des interactions moléculaires ou GeneOntology pour décrire les fonctions des gènes et d'autres composants biologiques. BioPAX et les ontologies utilisent tous deux un format orienté graphe. Les graphes constituent un moyen mathématique et informatique de représenter les interactions entre les entités, en décrivant les relations entre les entités (telles que l'ADN, l'ARN, les protéines et les métabolites) par le biais d'un ensemble de nœuds reliés par des arêtes.

Objectifs

Les objectifs de recherche de cette thèse englobent deux principaux aspects. Le premier consiste à développer une approche intégrative pour traiter des données hétérogènes à la fois issues de sources expérimentales et de bases de données publiques. Cette approche utilise le formalisme BioPAX et la théorie des graphes pour organiser et analyser les données à travers plusieurs couches. Puis, en cas d'étude, notre objectif est d'étudier le concept d'efficience alimentaire chez les porcs, un phénotype clé dans le contexte de la production durable de viande. Les résul-

tats mettent en lumière les mécanismes biologiques et les acteurs régulateurs qui influencent ce phénotype et contribuent à une meilleure compréhension de l'efficacité alimentaire.

Contributions

Traitement de la non-conformité des données et de la redondance : Méthodes de détection et de correction pour améliorer l'intégrité des données

L'utilisation d'ontologies et de formats normalisés facilite l'interopérabilité, un aspect crucial de l'ingénierie des données biologiques. Cependant, cela ne résout pas complètement tous les défis liés à l'intégration de données de types et de sources hétérogènes. Pour résoudre les problèmes de non-conformité et de redondance dans les bases de données, j'ai contribué au développement de méthodes, indépendantes des données, visant à détecter et à corriger de tels problèmes dans des formats standards. Les formats RDF de type graphe et les formats basés sur OWL offrent la possibilité d'être interrogés à l'aide d'outils du web sémantique tels que le langage de requête SPARQL. Les bases de données utilisant le format BioPAX présentent fréquemment des complexes moléculaires redondants, avec des propriétés identiques mais des identifiants distincts. De plus, ces bases de données contiennent souvent des complexes non valides, dans lesquels les composants eux-mêmes sont des complexes, aboutissant à une représentation récursive qui diffère de la représentation plane exigée par les spécifications du format. En conséquence, une telle non-conformité et cette redondance introduisent des modifications dans le graphe, impactant les analyses ultérieures. Premièrement, elles introduisent des problèmes de généricité, car les représentations redondantes masquent des redondances implicites, conduisant à des complications dans l'analyse des données. Deuxièmement, ces complexes existent dans diverses représentations sémantiques qui ne correspondent pas nécessairement aux interprétations biologiques, aggravant davantage le problème de la redondance. Enfin, ces structures modifient artificiellement la topologie du graphe, en augmentant la longueur du chemin entre les nœuds du graphe et compromettent l'analyse du réseau d'interaction.

Dans le cadre de cette recherche, visant à résoudre les problèmes de non-conformité et de redondance dans les bases de données pour faciliter leur analyse, un travail supplémentaire a également conduit à une publication dans *Bioinformatics*. Cet article traite de la différenciation entre la reproductibilité et la redondance implicite au sein des bases de données d'interactions protéines-protéines, et est disponible en annexe.

Combinaison de données expérimentales et de bases de connaissances pour concilier les niveaux moléculaire et cellulaire

Après les corrections nécessaires, l'utilisation de formats normalisés et d'ontologies offre le potentiel d'intégrer efficacement des données multimodales et multi-sources. Cependant, pour intégrer des données expérimentales avec des connaissances publiques disponibles dans des bases de données à différentes échelles, nous devons interroger plusieurs sources de données. Pour relever ce défi, j'ai développé une méthode basée sur des requêtes fédérées. Cette approche permet d'interroger simultanément plusieurs sources de données, facilitant ainsi l'intégration d'informations provenant de diverses origines. En tirant parti de listes expérimentales de transcrits de gènes et de métabolites présentant des co-variations, l'approche consiste à interroger un graphe métabolique pour identifier les nœuds annotés avec les identifiants d'intérêt. Cette étape facilite la mise en place de connexions entre plusieurs niveaux -omiques par le biais d'interactions, assurant une prise en compte complète des connaissances du système dans l'analyse.

Analyse des données et extraction de connaissances à l'aide de métriques basées sur les graphes

Dans le graphe métabolique enrichi, une analyse de parcours de graphe est effectuée pour déterminer les schémas d'organisation et les relations entre les entités d'intérêt. Cette étape vise à fournir une compréhension approfondie de la manière dont ces entités sont interconnectées et de leur contribution au phénotype étudié. L'accent est mis sur l'évaluation des chemins qui relient les entités au sein du graphe. De plus, l'analyse de parcours de graphe examine les interactions biochimiques traversées, facilitant les transitions entre les entités. En examinant ces interactions, nous pouvons identifier des molécules, des événements moléculaires et des processus potentiels qui pourraient réguler ou contribuer au phénotype observé. Pour faciliter l'analyse du graphe enrichi, le système de gestion de base de données (SGBD) de graphe Neo4j est utilisé. Ce système de gestion de base de données NoSQL est spécifiquement conçu pour manipuler efficacement les données des graphes. Il offre des fonctionnalités de stockage, de requête, et permet des opérations complexes de traversée de graphes et de correspondance de motifs, permettant une exploration et une analyse efficaces des entités biologiques interconnectées.

Structure de la thèse

Après cette introduction, la thèse se compose de 6 chapitres supplémentaires :

— **Chapitre 2** : Contexte. Ce chapitre présente des concepts importants et l'état de l'art en biologie des systèmes, en méthodes d'intégration multi-omiques, en Web Sémantique et en efficacité alimentaire. Son objectif est de fournir une compréhension du contexte de cette thèse et de la position de l'approche proposée et adoptée dans ce travail.

— **Chapitre 3** : Bases de données de connaissances et données expérimentales. Le but de ce chapitre est de présenter les données expérimentales, les bases de données et les ontologies utilisées dans ce travail.

— **Chapitre 4** : Correction des complexes moléculaires dans les normes BioPAX pour enrichir les interactions et détecter les redondances en utilisant les technologies du Web Sémantique. Ce chapitre est constitué de notre article publié, détaillant une méthode pour résoudre les problèmes de non-conformité prévalents dans les bases de données stockées au format BioPAX. La méthodologie améliore non seulement la conformité de ces bases de données, mais automatise également l'analyse du graphe en rectifiant la topologie des complexes.

— **Chapitre 5** : Petits réseaux de gènes exprimés dans le sang total et relations avec les profils de métabolites circulants fournissent des informations sur la variabilité interindividuelle de l'efficacité alimentaire chez les porcs en croissance. Dans ce chapitre, nous élucidons le processus d'identification de modules de gènes co-exprimés associés à l'efficacité alimentaire et de leurs liens avec les métabolites et les concentrations d'acides gras. Notre étude établit un lien entre les données de transcriptome et de métabolome, révélant des connexions entre l'immunité et la composition en acides gras.

— **Chapitre 6** : Une approche basée sur les graphes pour identifier les connexions complexes dans des réseaux biologiques hétérogènes. Ce chapitre présente une méthode d'intégration de données multi-omiques à l'aide des technologies du Web Sémantique. Notre méthode d'analyse, appliquée sur l'exportation de la base de données Reactome en BioPAX, identifie des chaînes de relations, basées sur la connaissance, entre des nœuds statistiquement liés dans des jeux de données biologiques, facilitant l'explication et suggérant des acteurs biologiques.

— **Chapitre 7** : Discussion. Ce chapitre met en évidence nos contributions et discute des perspectives d'amélioration de la méthode et des orientations futures de recherche.

SCIENTIFIC PRODUCTION

JOURNAL PAPERS

- ⁎ **Camille Juigné**, Emmanuelle Becker and Florence Gondret. "Small networks of expressed genes in the whole blood and relationships to profiles in circulating metabolites provide insights in inter-individual variability of feed efficiency in growing pigs". *BMC Genomics* 24, 647 (2023) https://doi.org/10.1186/s12864-023-09751-1. **(chapter 5 of the dissertation)**
- ⁎ **Camille Juigné**, Olivier Dameron, François Moreews, Florence Gondret, Emmanuelle Becker. "Fixing molecular complexes in BioPAX standards to enrich interactions and detect redundancies using Semantic Web Technologies". *Bioinformatics*, 39-5 (2023) https://doi.org/10.1093/bioinformatics/btad257. **(chapter 4 of the dissertation)**
- ⁎ Marc Melkonian, **Camille Juigné**, Olivier Dameron, Gwenaël Rabut, Emmanuelle Becker. "Towards a reproducible interactome: semantic-based detection of redundancies to unify protein-protein interaction databases". *Bioinformatics*, 38-6 (2022) https://doi.org/10.1093/bioinformatics/btac013.

SUBMITTED JOURNAL PAPERS

- ⁎ **Camille Juigné**, Emmanuelle Becker, Océane Carpentier, Florence Gondret and Olivier Dameron. "A graph-based approach to identify complex connections in heterogeneous biological networks". To be submitted in *Bioinformatics*. **(chapter 6 of the dissertation)**

PEER-REVIEWED CONFERENCES PROCEEDINGS

- ⁎ **Camille Juigné**, Olivier Dameron, Florence Gondret, Emmanuelle Becker. "A method to identify target molecules and extract the corresponding graph of interactions in BioPAX". *BBCC2022 - Bioinformatics and Computational Biology Conference, Dec 2022, Virtual, Italy* - Oral presentation. https://hal.science/hal-03876091

- ⁎ **Camille Juigné**, Olivier Dameron, François Moreews, Florence Gondret, Emmanuelle

Becker. "Detection and correction of non-conformities and redundancies in complexes of molecules in BioPAX." *Journées Ouvertes en Biologie, Informatique et Mathématiques (JOBIM), Jul 2022, Rennes, France* - Oral presentation. https://hal.science/hal-03752473

CONFERENCE ABSTRACTS

* **Camille Juigné**, Emmanuelle Becker, Florence Gondret. "Combined transcriptomics and metabolomics in the whole blood to depict feed efficiency in pigs." *European Federation of Animal Science (EAAP2023)* - Oral presentation. https://hal.science/hal-04103974

TALKS AND POSTERS

* **Camille Juigné**. "Integration and analysis of heterogeneous biological data modelled with multilayer graphs for a better understanding of feed efficiency". *Séminaire DIGIT-BIO INRAE, Dec 2022, Ecully, France* - Poster. https://hal.inria.fr/hal-03880428

* **Camille Juigné**. "Integration and analysis of heterogeneous biological data modelled with multilayer networks and applied for a better understanding of variations in feed efficiency". *Journées Numériques et Environnement de l'Inria, Oct 2021, Rennes, France* - Oral presentation.

* **Camille Juigné**. "Integration and analysis of heterogeneous biological data modelled with multilayer networks and applied for a better understanding of variations in feed efficiency". *Journées des doctorants de l'unité PEGASE de l'INRAE, Apr 2022, Saint-Gilles, France* - Oral presentation.

* **Camille Juigné**, Olivier Dameron, François Moreews, Florence Gondret, Emmanuelle Becker. "Detection and correction of non-conformities and redundancies in complexes of molecules in BioPAX." *Journées scientifiques de l'Ecole Doctorale EGAAL, Jul 2022, Rennes, France* - Oral presentation.

* **Camille Juigné**. "Considering molecular complexes for exhaustively connecting transcriptome and metabolome in BioPAX." *Journées du département Data Knowledge Managment de l'IRISA, Feb 2022, Rennes, France* - Poster.

SCIENCE POPULARIZATION

- **Camille Juigné**. La mélodie de la vie. Rédaction d'un texte de vulgarisation scientifique pour le pôle médiation scientifique de l'Inria, illustré avec un illustrateur professionnel, 2022. https://project.inria.fr/matheseunesacreehistoire/news/

TABLE OF CONTENTS

Acknowledgment, remerciements **3**

Résumé en français **7**

Scientific Production **13**

1 Introduction **23**
1.1 Context 23
1.2 Objectives 24
1.3 Contributions 25
1.3.1 Addressing data non-conformity and redundancy: Detection and correction methods for enhanced data integrity 25
1.3.2 Combining experimental data and knowledge bases to reconcile molecular and cellular levels 25
1.3.3 Data analysis and knowledge extraction using graph-based metrics . . . 26
1.4 Outline of the dissertation 26

2 Background **29**
2.1 Unimodal Omics Data 29
2.1.1 Defining the biological data 30
2.1.1.1 Genomics 30
2.1.1.2 Transcriptomics 31
2.1.1.3 Proteomics 33
2.1.1.4 Metabolomics 34
2.1.2 Storage : specialized databases and ontologies 35
2.2 Multimodal Omic Data Analysis 36
2.2.1 Benefits of multiomics 36
2.2.2 Existing approaches for multi-omics integration 37
2.2.2.1 Statistical integration and dimensionality reduction 38
2.2.2.2 Network-based integration and active modules research . . . 40

Networks in biology . . . 40
Integration approaches . . . 42
Identifying relevant modules . . . 43
2.3 Knowledge integration and representation . . . 45
2.3.1 Semantic Web, RDF Knowledge graphs and SPARQL . . . 45
2.3.2 Graph databases: Neo4j, Cypher and Neosemantics . . . 48
2.4 Case study: networks of entities associated with variability in feed efficiency of growing pigs . . . 49
2.4.1 Context . . . 49
2.4.2 Definition . . . 50

3 Knowledge databases and experimental data 53
3.1 Databases . . . 53
3.1.1 UniProtKB . . . 53
3.1.2 ChEBI . . . 55
3.1.3 Reactome . . . 55
3.2 Experimental data . . . 57

4 Fixing molecular complexes in BioPAX standards 61
4.1 Abstract . . . 61
4.2 Introduction . . . 62
4.2.1 Molecular complexes and biological interactions in system biology . . . 62
4.2.2 Description of complexes in `BioPAX`. . . . 63
4.2.3 The Reactome use-case. . . . 63
4.2.4 Motivations and results . . . 64
4.3 Approach . . . 64
4.3.1 Definition of invalid recursive complexes . . . 64
4.3.2 Invalid recursive complexes in interactions . . . 64
4.3.3 Redundancies . . . 65
4.4 Methods . . . 65
4.4.1 Identifying invalid complexes . . . 65
4.4.2 Fixing the invalid complexes . . . 66
4.4.3 Identifying redundant complexes . . . 68
4.5 Results . . . 68

4.5.1 Invalid complexes represent a significant part of complexes in the BioPAX description of Reactome . . . 68
4.5.2 Fixing invalid complexes increases the average number of components participating to complexes in Reactome . . . 68
4.5.3 Fixing invalid complexes reduces the path length from a complex to each of its components . . . 69
4.5.4 Fixing invalid complexes improves the detection of redundant complexes 70
4.5.5 Application to non-Human organisms in the Reactome database . . . 72
4.6 Discussion and perspectives . . . 72

5 Small networks of expressed genes in blood and relationships to metabolic profiles 75
5.1 Abstract . . . 75
5.2 Introduction . . . 76
5.3 Material and methods . . . 78
5.3.1 Ethics . . . 78
5.3.2 Origin of phenotypic data . . . 78
5.3.3 Transcriptomic dataset . . . 79
5.3.4 Metabolomic dataset . . . 79
5.3.5 Construction of the weighted gene co-expression networks . . . 80
5.3.6 Detection of modules of co-expressed probes and their relationships with animal phenotypic traits . . . 81
5.3.7 Biological functional enrichment in modules of co-expressed probes . . 82
5.3.8 Establishing profiles of circulating metabolites and evaluating connections between metabolic and transcriptomic levels . . . 83
5.4 Results . . . 83
5.4.1 Definition of gene co-expression network in the whole blood of pigs . . 84
5.4.2 Relationships between modules of co-expressed genes and animal phenotypic traits . . . 84
5.4.3 Close-vicinity of the different modules of co-expressed genes . . . 87
5.4.4 Functional enrichment of the modules in biological processes . . . 87
5.4.5 Hierarchy of expressed genes in the modules related to feed efficiency traits . . . 91
5.4.6 Metabolic profiles in the whole blood . . . 92
5.4.7 Connecting the two omics levels . . . 96

5.5 Discussion . . . 98
5.5.1 Analyzing inter-individual variability in feed efficiency . . . 98
5.5.2 Enriched pathways in co-expressed genes modules related to variability in feed efficiency . . . 99
5.5.3 Important genes in molecular networks related to feed efficiency . . . 101
5.5.4 Relationships between transcriptomic and metabolic levels in the definition of feed efficiency or related traits . . . 102

6 A graph-based approach to identify connections in heterogeneous biological networks 107
6.1 Introduction . . . 108
6.2 Background . . . 109
6.2.1 Databases and ontologies in biology . . . 109
6.2.1.1 The UniProt database . . . 109
6.2.1.2 The ChEBI Ontology . . . 109
6.2.1.3 The BioPAX ontology . . . 110
6.2.1.4 The Reactome knwoledgebase . . . 111
6.2.2 Weighted co-expressed gene networks coupled to metabolic profiles as a use-case . . . 111
6.2.2.1 Weighted co-expressed gene networks involved in feed efficiency . . . 111
6.2.2.2 Integration of transcriptomic and metabolic -omics levels on the same Reactome graph to explore connections . . . 112
6.3 Methods . . . 112
6.3.1 Retrieving `Gene` and `Protein` in the BioPAX export of Reactome using federated SPARQL queries . . . 113
6.3.2 Retrieving `SmallMolecules` in the BioPAX export of Reactome using federated SPARQL queries . . . 114
6.3.3 Computing paths between nodes of interest in Neo4j using Cypher queries 114
6.3.4 Biochemical reaction cascades and their regulation in modules of co-expressed genes and comparison with randomizations . . . 115
6.4 Results . . . 116
6.4.1 Proteins retrieved by their UniProt ID in Reactome . . . 116
6.4.2 SmallMolecule retrieved by their ChEBI ID in Reactome . . . 116

6.4.3 Graph traversal and paths connecting molecules of interest 117
6.5 Discussion 120

7 Discussion **123**
7.1 Benefits of our approaches 124
7.2 Limitations of the study 127
7.3 Perspectives and potential future improvement and research directions 129
7.4 Conclusion 131

Bibliography **133**

CHAPTER 1

INTRODUCTION

1.1 Context

Recent technological advances in biological data acquisition result in an exponential growth of multi-modal (different types) data generated across various laboratories, experiments and conditions. This rapidly expanding data volume is causing scientific challenges in terms of storage, standardization and analysis [1, 2].

This thesis specifically addresses the last two challenges, aiming to maximize the utilization of data from experiments and leveraging the knowledge available in various biological databases and ontologies. The objective is to gain new insights from the extensive data analysis, considering the context of the knowledge stored in databases and ontologies, in order to enhance our depiction and understanding of systems biology.

We hypothesized that adopting a holistic perspective on biological organization can lead to a better understanding of complex processes by representating the numerous relationships between different biological entities.

This implies taking advantage of various -ome modalities, such as the genome, transcriptome, proteome, metabolome, etc., by establishing connections between them and encompassing some bias (selectivity, sensitivity, etc.) associated with the data acquisition technologies for these diverse entity types. The field of -omics (i.e., technologies to acquire data related to the different -ome levels) encompasses a wide range of objectives, including understanding, preventing, diagnosing, and managing diseases and optimizing performance.

Most of the time, measurable changes in highly complex biological mixtures (cells, tissues, fluids) across experimental or environmental conditions are analyzed within a given level of biological organization (unimodal) using univariate or multivariate statistical methods, and are compared with existing knowledge through specialized expertise. Methods for the joint analysis of multimodal -omics data have also been developed to facilitate the mapping of altered pathways. These methods mainly focus on reducing data dimensionality by identifying important shared features and facilitating data manipulation. However, these approaches often overlook

the available knowledge regarding the physical relationships between molecules (activation, inhibition, complexes formation, interaction, etc.). This oversight may result in the loss of valuable information and the missing of essential insights or interdependencies. To address these challenges, we propose a comprehensive and systemic integration approach for the analysis of multimodal -omics data that takes into account the knowledge present in databases and knowledge bases.

There are a multitude of specialized databases, each focusing on specific types of data, employing different formats and having its own curation processes. As a result these databases are complementary and integrating them remains an important challenge due to their heterogeneity and the existence of potential overlaps between them. Notably, there are several databases dedicated to describing metabolic pathways, which facilitate the linkage of biological entities across different -omics layers. To address this integration challenge, the Biological Pathway Exchange (BioPAX) standard has emerged as a valuable solution. This format provides a comprehensive representation with great finesse of the knowledge on biological processes at the molecular and cellular levels. BioPAX enables the integration of diverse knowledge bases related to metabolic pathways and facilitates cross-referencing with external resources, including well-established ontologies such as UniProt for proteins, ChEBI for small molecules, MI for annotating molecular interactions or GeneOntology for describing the functions of genes and other biological components. Both the BioPAX and ontologies employ a graph-oriented format.Graphs provide a mathematical and informatics means to represent interactions between entities, capturing relationships between entities (such as DNA, RNA, proteins and metabolites) through a set of nodes connected by edges.

1.2 Objectives

The research objectives of this thesis encompass two main aspects. The first one is to develop an integrative approach to handle heterogeneous data from both experimental sources and public databases. This approach employs the BioPAX formalism and graph theory to organize and analyze data across multiple layers. As a use-case, our aim is to investigate the concept of feed efficiency in pigs, a pivotal phenotype in the context of sustainable meat production. Through the developed integrative approach, we seek to explore the underlying processes that contribute to feed efficiency. The results shed light on the biological mechanisms and regulatory actors influencing this phenotype and contribute to a deeper understanding of feed efficiency.

1.3 Contributions

1.3.1 Addressing data non-conformity and redundancy: Detection and correction methods for enhanced data integrity

The utilization of ontologies and standardized formats facilitates interoperability, a crucial aspect of biological data engineering. However, it does not completely address all challenges associated with integrating heterogeneous data types and sources. To address the issues of non-conformity and redundancy within databases, I have contributed to the development of data-independent methods aimed at detecting and fixing such problems in standard formats. Graph-like RDF and OWL-based formats offer the capability to be queried using semantic web tools like the SPARQL Protocol and RDF Query Language (SPARQL). Indeed, databases utilizing the BioPAX format frequently exhibit redundant molecular complexes with identical properties but distinct identifiers. Furthermore, these databases often contain invalid complexes, in which the components themselves are complexes, resulting in a recursive representation that differs from the flat representation required by the format specifications. Consequently, such non-conformity and redundancy introduce modifications within the graph, which impact the subsequent analyses. Firstly, they introduce genericity problems, as redundant representations mask implicit redundancies, leading to complications in data analysis. Secondly, these complexes exist in various semantic representations that may not necessarily align with biological interpretations, further exacerbating the issue of redundancy. Lastly, these structures artificially modify the topology of the graph, increasing the path length between graph nodes and compromising the analysis of the interaction network.

As part of this research focus aimed at addressing issues of non-conformities and redundancies in databases to facilitate their analysis, additional work has also led to a publication in *Bioinformatics*. This article deals with differentiating reproducibility and implicit redundancy within protein-protein databases, and is available in the appendix.

1.3.2 Combining experimental data and knowledge bases to reconcile molecular and cellular levels

After necessary corrections, the utilization of standardized formats and ontologies holds the potential for effectively integrating multi-modal and multi-sources data. However, to integrate experimental data with public knowledge available in databases across different scales, we need

to query multiple data sources. To address this challenge, I developed a method centered around federated queries. This approach enables simultaneous querying multiple data sources, facilitating the integration of information from various origins. Taking advantage of experimental lists of gene transcripts and of metabolites that exhibit co-variations, the approach consists in querying a metabolic graph to identify nodes annotated with the identifiers of interest. This step facilitates the establishment of connections between multiple -omics levels through interactions, ensuring comprehensive consideration of system knowledge into the analysis.

1.3.3 Data analysis and knowledge extraction using graph-based metrics

Within the enriched metabolic graph, a graph traversal analysis is performed to elucidate organizational patterns and relationships between the entities of interest. This step aims to provide a deeper understanding of how these entities are interconnected and how they contribute to the studied phenotype. The primary focus is on evaluating the paths that connect the entities within the graph. Furthermore, the graph traversal analysis examines the biochemical interactions that are traversed and then facilitates transitions between the entities. By investigating these interactions, we can identify potential molecules, molecular events and processes that might regulate or contribute to the observed phenotype. To facilitate the analysis of the enriched graph, the Neo4j graph database management system (DBMS) is employed. This NoSQL DBMS is specifically designed to handle graph data efficiently and effectively. It offers capabilities for storing, querying, and allows for complex graph traversals and pattern matching operations, enabling effective exploration and analysis of interconnected biological entities.

1.4 Outline of the dissertation

Following this Introduction, the dissertation consists of 6 more chapters:

— **Chapter 2**: Background. This chapter presents important concepts and the state of the art in systems biology, multi-omics integration methods, Semantic Web, and feed efficiency. It aims to provide an understanding of this dissertation's context and the position of the approach advocated and adopted in this work.

— **Chapter 3**: Knowledge databases and experimental data. The purpose of this chapter is to present the experimental data, databases, and ontologies used in this work.

— **Chapter 4**: Fixing molecular complexes in BioPAX standards to enrich interactions and detect redundancies using Semantic Web technologies. This chapter is made of our published article, detailing a method to tackle non-compliance issues prevalent in databases stored in the BioPAX format. The methodology not only enhances the conformity of these databases but also automates the analysis of the graph by rectifying the topology of the complexes within.

— **Chapter 5**: Small networks of expressed genes in the whole blood and relationships to profiles in circulating metabolites provide insights in inter-individual variability of feed efficiency in growing pigs. In this chapter, we elucidate the process of identifying modules of co-expressed genes associated to feed efficiency and their connections with metabolites and fatty acids concentrations. Our study establishes a link between transcriptome and metabolome data, revealing connections between immunity and fatty acid composition.

— **Chapter 6**: A graph-based approach to identify complex connections in heterogeneous biological networks. This chapter presents method to integrate multimodal -omics data using Semantic Web technologies. Our analysis method based on Reactome BioPAX export identifies knowledge-based chains of relationships between statistically related nodes in biological datasets, facilitating explainability and suggesting modulatory biological actors.

— **Chapter 7**: Discussion. This chapter highlights our contributions and discusses perspectives for enhancing the method and future research directions.

CHAPTER 2

BACKGROUND

2.1 Unimodal Omics Data

The field of biological data acquisition has experienced significant advancements in recent decades (Figure 2.1). These advancements in equipment and methodologies have revolutionized the way biological data are generated, enabling high-throughput and cost-effective acquisition for various biological specimens. Techniques such as micro-arrays and next-generation sequencing, chromatography and non-targeted mass spectrometry (MS) as well as Nuclear Magnetic Resonance (NMR) spectroscopy have significantly increased the speed and efficiency of data acquisition [3]. The technologies go well beyond the scope of standard chemistry techniques since they are capable of precise analyses of hundreds to thousands of molecules.

This technical evolution has given rise to the field of high-throughput -omics data analysis, which encompasses data obtained from various levels of life organisation, namely the genome, transcriptome, proteome, metabolome, lipidome, as well as the description of the microbiome. These omics data are characterized by their heterogeneity.

1. **Heterogeneity of entity type**: They represent distinct biological entities such as genes, transcripts, proteins, lipids, metabolites, each with its own unique chemical and physical characteristics. Some of them can be even separated according to their chemical characteristics, such as phosphorylated proteins (for the phospho-proteome).
2. **Heterogeneity of data**: There is heterogeneity in terms of the nature of the data itself, for instance, ranging from textual data for DNA or protein sequences, to binary data indicating presence or absence of the entities or quantitative data representing the abundance of molecules and phenotypic signatures, and qualitative data describing biological functions or interactions.
3. **Technical heterogeneity**: Finally, there is technical heterogeneity, which refers to variations in measurement techniques, experimental protocols, and data formats.

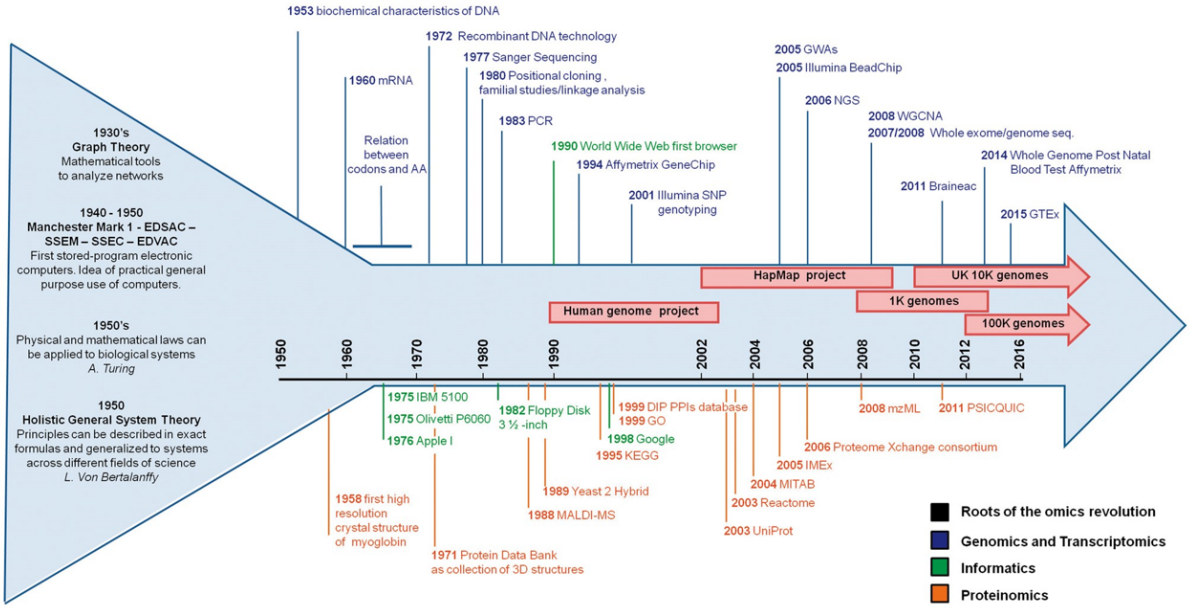

Figure 2.1 – A schematic chronological timeline of some of the key technological and theoretical advances in genomics, transcriptomics and proteomics leading to multiomics. It encompasses milestones such as major advances in graph theory in the 1930s, the development of DNA sequencing methods, and the advent of high-throughput technologies on RNA and proteins like next-generation sequencing and mass spectrometry, respectively, in the 2000s. It also acknowledges the creation of specialized knowledge repositories like UniProt and Reactome. (Figure extracted from *"The rise of omics data and their integration in biomedical sciences"* by Manzoni et al. [4]).

2.1.1 Defining the biological data

Before delving into the intricacies of analysis and management related to biological data, let us establish the fundamental concepts and connections among various biological entities.

2.1.1.1 Genomics

Genomics is the field of research that focuses on the study of genomes, including their sequence, structure, function, and evolution. The genome serves as the repository of an organism's genetic information and is composed of deoxyribonucleic acid (DNA), arranged into one or multiple chromosomes, depending on the species. These chromosomes are located in the cell nucleus and contain genes, which are specific segments of DNA that encode various molecules such as proteins, miRNAs and lncRNAs. DNA possesses a double helix structure consisting of two strands, with each strand composing of four fundamental building blocks called nucleotides or bases. These bases are represented by single letters: A for Adenine, T for Thymine, C for Cy-

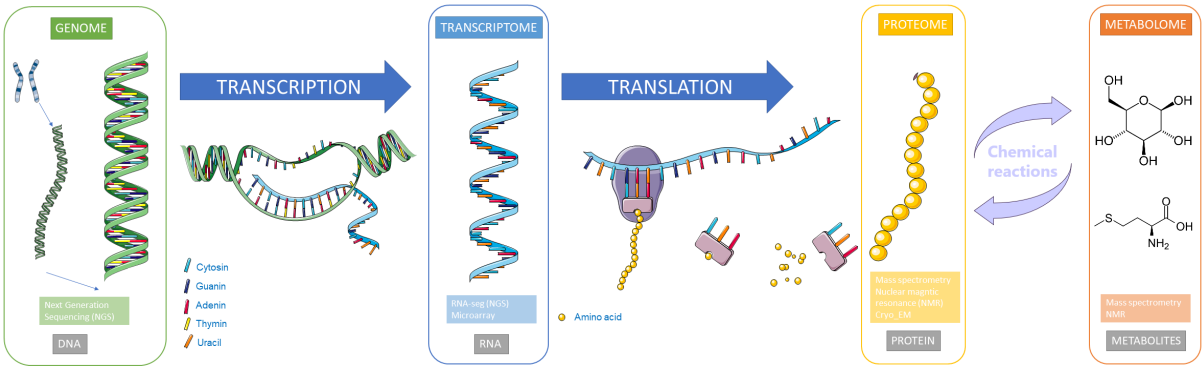

Figure 2.2 – Overview of the various -omes and their associated -omics technologies, highlighting their intricate relationships.
Parts of the figure were drawn by using pictures from Servier Medical Art. Servier Medical Art by Servier is licensed under a Creative Commons Attribution 3.0 Unported License (https://creativecommons.org/licenses/by/3.0/).

tosine, and G for Guanine. The genome sequence represents therefore a list of nucleotides comprising all the chromosomes. DNA sequencing is the process used to determine the sequence of a DNA fragment. Because the two stands of DNA are complementary, the measurement unit used for sequences is base pairs (pb). For example, the obtained sequence data for the sequencing of *Homo sapiens* represent 3.1 gigabases, 4.6 megabases for *Escherichia coli* and 2.7 gigabases for *Sus scrofa*. Both the human and pig genomes exhibit a similar order of magnitude in terms of number of genes, with slightly more than 20,000 genes. However, it's important to emphasize that a significant portion of the genomes does not encode proteins. This non-coding portion of the genome plays essential roles in regulating gene expression, influencing cellular processes, and contributing to the complexity of biological functions.

2.1.1.2 Transcriptomics

The process of converting the genetic information encoded in a gene into a functional molecule is known as gene expression. During gene expression, a specific strand of the DNA corresponding to the gene is utilized as a template to produce a complementary ribonucleic acid (RNA) strand by RNA polymerases, in a process called transcription. Some of the resulting RNA molecules, messenger RNA (mRNA) or transcripts, carry the coding sequences that guide the synthesis of proteins in a subsequent process known as translation. Although all cells in an organism possess the same genetic information in the form of DNA, the gene expression can vary significantly depending on various factors, such as the cell type, the developmental stage of the organism or the environmental conditions. Gene expression regulation stands as a fun-

damental mechanism, enabling cellular differentiation, morphogenesis, and the adaptability of living organisms to their environments. The expression level of a gene is directly correlated with the abundance of its mRNA copies within the cell. As the number of mRNA copies increases, the gene's expression rises proportionally.

Measuring the abundance of mRNA or studying gene expression levels is referred to as transcriptomics. The main techniques for screening the transcriptome are microarrays gene expression and RNA sequencing (RNA-seq).

Microarrays workflows involve the conversion of RNA copies into complementary DNA (cDNA) molecules through reverse transcription. These cDNA molecules can then interact with specific probes on the microarray that are designed to target particular genes of interest. When the sample containing RNA from these genes is exposed to the microarray, the cDNA molecules in the sample bind to their complementary probes on the microarray. It leads to the emission of fluorescent signals. Each signal's intensity is proportional to the number of copies of the specific gene in the sample. Microarrays relies on prior knowledge of the DNA sequences of the gene whose expression is being measured.

RNA-seq sequences the complementary DNA (cDNA) fragments that are generated from the RNA in the sample. The library preparation is optimized for enriching coding mRNAs. The obtained reads are aligned with a reference genome. This process allows for the quantification of gene expression, providing information about which genes are expressed and at what levels in the sample. RNA-seq offers several advantages over microarrays, including the ability to detect and quantify transcripts without the need for pre-designed probes. RNA-seq can provide a more comprehensive and unbiased view of the transcriptome, by identifying more differentially-expressed protein-coding genes and providing a wider quantitative range of expression level changes compared to microarrays. Additionally, RNA-seq can detect splice junctions, gene fusions, and single-nucleotide polymorphisms (SNPs). This makes it a preferred choice for many modern transcriptomic studies but it remains more costly than microarrays and requires more extensive reference data to fully leverage additional RNA-Seq data, especially for non-coding sequences [4]. Typically, around 80% of the differentially-expressed genes identified with microarrays overlap with RNA-seq data [5].

At the conclusion of these analyses, the result is a table of numerical data representing probe expression levels (for microarrays) or the number of reads for each sequence (for RNA-seq). The corresponding gene names can be added based on the microarray annotation or by mapping the sequenced reads to known genes in the case of RNA-seq.

2.1.1.3 Proteomics

Proteome represents the set of proteins in a given sample at a specific time. During translation, single-stranded mRNA is read to synthesize proteins. mRNA is composed of nucleotides A, U, C and G, and each triplet (i.e. codon) of these nucleotides corresponds to one of the 20 amino acids that are to be synthesized. This process of reading the mRNA codons and assembling the corresponding amino acids leads to the formation of the complete protein. Proteins exhibit various lengths and configurations, and they can go through chemical modifications to reach their functionality. They can be further modified by phosphorylation, glycosylation, etc. these extensive post-translational modifications along the production pathways leading to biologically active proteins. Also, a single protein may exist in multiple forms, known as isoforms, which result from alternative splicing processes. Proteins play a diverse and essential role, serving as key components responsible for numerous cellular activities. To execute various and fundamental biological functions, proteins engage in interactions through physical contacts, often forming protein complexes where two or more proteins bind together. These interactions are referred to as protein-protein interactions (PPIs). From a computational perspective, protein-protein interactions are frequently studied using graph theory, creating what is known as the interactome. This approach helps unravel the complex network of interactions between proteins and sheds light on their functional relationships.Proteomics encompasses thus a wide and intricate spectrum of research, spanning from the investigation of protein structures and folding [6–8], to the exploration of proteins interactions [9], and the study of their functions which are intricately linked to their folding and complex assembly.

Experimental techniques for studying proteins are equally diverse. For example, mass spectrometry is employed to determine protein mass and amino acid composition, nuclear magnetic resonance (NMR), X-ray crystallography and cryogenic electron microscopy (cryo-EM) are used to investigate protein folding, while gel electrophoresis (Western blot) is employed to identify specific proteins (antigen-antibodies) in samples. To identify PPIs, methods such as double hybrid and affinity purification coupled with mass spectrometry, as well as other luminescence-based methods, are utilized.

Depending on the chosen perspective to study the proteome, the data are considered as quantitative, semi-quantitative or only refer to presence or the absence of specific proteins in the biological samples. They can take the form of text for protein sequences or exist in network formats (interactome) for protein-protein interactions (PPIs), and they can vary in types for other studies.

2.1.1.4 Metabolomics

Metabolomics focuses on the detection, identification, and quantification of metabolites (small molecules) from complex biological matrices. Metabolites are small molecules produced and transformed in cellular metabolic processes, including amino acids, glucose, lipids, and various other molecules. To determine which metabolites are present in a sample and in what quantities, methods such as nuclear magnetic resonance (NMR) and Liquid Chromatography coupled to tandem Mass Spectrometry (LC-MS) are used. LC-MS and 1H-NMR are untargeted high-throughput methods that generate substantial amount of data. LC-MS faces the challenge of translating signals into metabolite identities because it can provide the atomic formula of the analytes. In contrast, 1H-NMR relies on well-annotated peaks in the spectrum obtained from structural moieties, enabling more precise biological interpretation [10] but is inherently less sensitive than LC-MS. LC-MS and 1H-NMR provide complementary data as LC-MS can identify certain functional groups (sulfate and nitro groups) which are 1H-NMR silent [**Gathungu2020**].

As a result, researchers typically obtain metabolic profiles that provide an overview of the types and concentrations of metabolites present in a specific biological sample at a given moment. The analysis of these metabolic profiles can help in differentiating two groups of individuals (e.g., healthy and diseased), identifying biomarkers (generally defined as the metabolites that account for most of the variation) associated with specific phenotypes or diseases, or assessing changes in the metabolic profiles after the application of a treatment (diet, medicine, etc.) [11]. Metabolomics is also used more broadly to understand biological mechanisms, elucidating metabolic pathways, enzymatic reactions and molecular interactions that occur within a biological system.

What we have seen in this section provides an overview of the four primary -omics levels. These, of course, are not the only ones, but they form a solid and substantial foundation for the field of multi-omics research. From these four presentations, we can observe that different areas of research in biology are inherently interconnected. Molecules are synthesized from other molecules, and other molecules regulate these processes. We have also highlighted the heterogeneity of these biological entities and the data describing them.

2.1.2 Storage : specialized databases and ontologies

Concurrently with the advancements in data acquisition methods, the creation of large-scale biological databases has become increasingly prevalent. These databases serve as repositories for storing and organizing vast amounts of biological data generated from various organisms and experiments. They play a crucial role in facilitating data sharing, integration, and analysis in the field of life sciences. Indeed, they stand as valuable resources for researchers, providing curated and comprehensive data that can be leveraged to gain insights into biological processes, drive scientific discoveries, and advance our understanding of the complexities of phenotypes.

Because of biochemical data heterogeneity, these biological databases encompass diverse types of data, including genomic sequences [12], protein sequence and functional information [13], gene expression profiles [14], protein structures [15], metabolic pathways [16], and clinical information [17].

To further enhance the value and interoperability of biological databases, initiatives based on Semantic Web have emerged. Web semantics is a field of computer science that focuses on the representation and organization of knowledge (see section 2.3). By applying semantic principles, such as the use of ontologies, these initiatives aim to improve data integration, interoperability, knowledge sharing and complex data reasoning. An ontology is a formal and explicit representation of knowledge within a particular domain. This representation includes clear and precise definition of entities or concepts, along with the relationships that describe how these different entities are related to each other. Each entity is also associated with properties describing its attributes and characteristics. In practical terms, within a given context that employs this data model (i.e. the given ontology), instances of class entities and relationships are created based on the defined data schema.

There are several prominent ontologies that play a crucial role in the field of life sciences. These ontologies serve as standardized vocabularies or knowledge frameworks that capture domain-specific concepts and relationships. They provide a common language for describing biological entities, processes, functions, and annotations, thereby promoting data integration and interoperability across different databases and resources. For example, the UniProt ontology provides a comprehensive and structured representation of protein-related information, including protein sequences, functions, and interactions [13]. The Chemical Entities of Biological Interest (ChEBI) ontology focuses on small molecules and their chemical properties, enabling the consistent annotation and classification of diverse compounds [18]. The Proteomics Standards Initiative Molecular Interaction (PSI-MI) ontology facilitates the representation and exchange of data related to molecular interactions [19]. The Gene Ontology (GO) ontology

captures biological components, processes, and functions, serving as a standardized vocabulary for annotating gene-related information [20]. The complexity increases when dealing with phenotypes. However, ongoing efforts are being made to establish unified semantic terms. For human phenotypes, the Human Phenotype Ontology (HPO) fills the role [21]. For production, health and environmental traits in livestock, ontologies like ATOL, AHOL, EOL are being developed [22]. One recent example of the application of web semantics in bioinformatics is the development of ontologies specifically intended for the COVID-19 domain, for instance the COVID-19 Ontology [23].

2.2 Multimodal Omic Data Analysis

2.2.1 Benefits of multiomics

Single-omic level analyses have undoubtedly played a pivotal role in advancing our comprehension of biological systems and continue to be valuable. Nevertheless, in some respects, there are growing challenges in further refining these analyses. Conversely, we are now generating an ever-increasing volume of multi-omics data for identical samples or identical scientific questions. This proliferation is evident when considering the scale of platforms such as Omics Discovery Index (OmicsDI) (3.4M datasets in September 2023), which archive publicly available omics datasets [24]. Hence, it is imperative to explore strategies for maximizing the utilization of the wealth of data generated on a daily basis. Considering the inherent connection of -omic levels and their mutual influence, embracing joint and multi-omic analyses has become a clear and valuable solution to extract more comprehensive insights from newly generated data as well as from the abundance of existing material.

The heterogeneous data produced at different organizational scales in organisms are inherently interdependent. Analyzing these datasets independently of each other can limit our understanding of the complex biological processes that underlie biological phenomenons. The integration of multi-modal data is essential for achieving a comprehensive and interconnected perspective of biological systems. Each -omics level provides valuable insights into specific aspects of cellular and organismal function and helps to understand how these different components interact and mutually impact one another.

However, integrating heterogeneous data is challenging due to the high heterogeneity of data in terms of technical, biological and semantic aspects [25]. A review by Eicher et al. emphasizes the particular difficulties associated with integrating metabolomic data with other omics

layers [26]. Indeed, the complication arises from the absence of a direct association between metabolites and transcripts, unlike transcriptomics and proteomics, where most transcripts can be mapped to a single protein, allowing for direct profile comparisons [27].

Nonetheless, addressing these challenges is achievable through various techniques that we introduced in the following section. Indeed, some multi-omics studies have demonstrated the value of this holistic approach, particularly in cancer research. For instance, multi-omic analyses have enable the identification of underlying molecular signatures across different -omic layers associated with specific cancer phenotypes, classification of cancer types, and assistance in clinical decision-making [28–31]. These studies have been made possible by major projects such as The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC). The TCGA project, in particular, has facilitated the generation of large datasets comprising heterogeneous and interconnected data on cancer, including tissue exome sequencing, copy number variations (CNV), DNA methylation, gene and microRNA expression, as well as relevant physiological and clinical data [32]. Moreover, studies that integrate various types of data have also emerged in other domains. For instance, research on the impact of environmental factors on different -omics layers, such as the work of Angione et al. on assessing the influence of various conditions on E. coli growth using a multilayer graph [33] or the work of Rattray et al. integrating genomics and metabolomics to determine the causes and effects of environmental exposures [34]. Additionally, there are investigations into different human phenotypes in relation to human microbiomes, enabling by the NIH Human Microbiome Project (iHMP) [35, 36].

2.2.2 Existing approaches for multi-omics integration

The number of reviews published on multiomics analysis tools in recent years demonstrates the expanding interest and importance of this field [26, 27, 30, 37–41]. These reviews highlight the advancements, methodologies, and applications of multi-omics analysis.

The analysis of multimodal data is guided by the nature of the data themselves and the specific research objectives. Integration methods, as described by Zhou et al., can be classified into two main approaches: statistical integration and network-based integration [38]. Statistical integration methods are based on identifying common patterns across diverse -omics datasets through dimension reduction techniques. These methods focus on uncovering shared features and reducing data complexity. In contrast, network-based integration methods adopt a holistic approach to understand biological systems. They use interconnected networks that represent

various biological entities, thereby identifying significant connections and relationships within these networks.

2.2.2.1 Statistical integration and dimensionality reduction

When working with data from various -omics analyses, the first challenge encountered is the dimensionality of the data. Indeed, in biology, the number of measured elements often highly exceeds the number of samples, which leads to unfavorable high-dimensional data. This phenomenon is commonly referred to as 'the curse of dimensionality' (Richard Bellman, 1961).

Principal Component Analysis (PCA) is a commonly used technique to address this issue. PCA reduces the dimensionality of the data by finding a new set of uncorrelated variables, known as principal components, that capture the maximum variance in the data. Each principal component is a linear combination of the initial variables, but principal components are orthogonal and therefore uncorrelated. PCA aims to put the maximum possible information in the first component, then the maximum remaining information in the second component and so on. These principal components are used in subsequent analyses in place of the numerous original variables.

However, when applying PCA to multimodal datasets, differences in variances between modalities can introduce bias in the results if combined. To address this challenge, several methods have been developed that consider the contributions of individual modalities during the dimension reduction process. They include Sparse Principal Component Analysis (sPCA) from the mixOmics R package [42]. sPCA is an extension of PCA that assigns weights to variables and penalizes the less informative ones, resulting in a subset of the most informative variables rather than a combination of all original variables. Among Multiple Factor Analysis-based approach, the padma method extends PCA to analyze multi-omic dataset, investigating biological variation at the pathway level by aggregating information across different sources. [43]. Partial Least Square (PLS) based integration methods have also proven valuable in multi-omics analyses and offering a complementary solution. Thanks to PLS these methods are designed to identify the variables among numerous variables that contribute significantly to the observed differences between samples, effectively reducing dataset dimensions while retaining the most pertinent information. PLS achieves this through an iterative process that constructs novel variables termed latent components which are linear combinations of the original variables. The constructions is carried out by maximizing covariance. PLS-based approaches can be employed for variable selection and performing classification to differentiate groups of individuals or phenotypes as depicted in PLS-DA [44, 45]. For instance, employing the sparse Multi-Block Partial

Least Squares (sMBPLs) method enables the detection of heterogeneous sets of gene expression regulators within multilayer datasets [46]. In a more recent context, the mixomics framework DIABLO [47] has emerged as a valuable tool for variable selection capitalizing on latent components to facilitate the discovery of biomarkers. The mixOmics group has also introduced the MINT tool, which relies on PLS for vertical integration. This approach involves utilizing identical variables obtained from different sources and emphasizing variations across tissues [48, 49].

Matrix factorization is another prominent method for achieving omics integration through dimension reduction [37]. It involves projecting the variables of a dataset onto a lower-dimensional space, thereby capturing the underlying patterns and relationships in the data [50]. Each -omics dataset is typically represented as an extensive matrix, which can be further decomposed into a product of two matrices. A crucial constraint in this decomposition is that one of the matrices in the product is shared across the various datasets. These methods take into account the specific characteristics and contributions of each modality when factoring the matrices, leading to more accurate and informative outcomes. For example, the MOFA tool [51] employs matrix factorization to reduce dimensionality and uses a Bayesian model to account for the complexity and uncertainty inherent in -omics data. MOFA's objective is to identify a restricted number of latent factors (hidden) that capture the shared source of variability across all input -omics datasets obtained from a common set of samples (horizontal integration). MOFA has evolved into MOFA+ to address various challenges [52]. This newer version is well-suited for handling large-scale datasets and offers the capability for both vertical integration (different samples) and horizontal integration (different modalities), providing a more comprehensive and flexible approach. The group have also more recently, developed the MEFISTO [53] specifically designed for processing temporal data using dimension reduction method.

Another possible approach is to use machine learning techniques. For instance, supervised variational autoencoders have been employed to integrate transcriptomic data with neuroimaging data to calculate disease progression scores [54]. Similarly, neural networks have been used to identify relationships between the microbiome and the metabolome providing insights into the causes of dysregulations in disease (MiMeNet) [55].

The approaches discussed here encompass multivariate analyses of data derived from diverse omics modalities. These methodologies have proven to be effective as they possess the capacity to manage extensive datasets by simplifying them, retaining only the information that accounts for the most significant variability in the outcomes. This capability facilitates the recognition of dependency relationships among variables, making them efficient for tasks

such as classification and prediction.

Despite their advantages, dimension reduction methods present some limitations, primarily in terms of interpretability. The transformed representation obtained by these techniques may not always be easily interpretable and determining the optimal choice of components can be challenging. Furthermore, these statistical associations often lack of biological explainability and are not supported by the available knowledge about the physical relations between molecules, including aspects like activation, inhibition, interaction, control or participation in a common reaction. Additionally, these methods typically impose a strong constraint that requires integrating data from the same study, either based on the same individuals (e.g., DIABLO, MOFA) or identical measurements (e.g., MINT). This constraint can is limiting the applicability of these methods when working with diverse datasets from various sources or studies.

Overcoming these limitations remains a subject of ongoing research in the field of multi-omics integration. Novel methods and techniques are being developed to address these challenges, seeking to enable more flexible, robust, and comprehensive integration of diverse datasets with varying characteristics and missing data.

2.2.2.2 Network-based integration and active modules research

Other methods focus on the interconnected and dependent nature of biological entities, with the aim of retaining almost all measured variables to prevent loss of important signals and to emphasize relationships across different -omics levels. This approach allows the identification of active modules of entities and is exemplified by network-based methods, whether they are driven by knowledge or not.

Network-based approaches are increasingly prevalent and well-suited to the nature of biological data, which is often manifold and interconnected. These approaches rely on fundamental graph theory principles, utilizing graphs to represent the data and the relationships that interconnect them. Entities are represented by vertices (or nodes), and existing relationships (interactions) between two nodes are represented by edges connecting the respective nodes. These edges can be either directional or non-directional, depending on whether the relationships between the nodes have a direction or not. They are referred to as directed or undirected graphs.

Networks in biology To begin this section, let us now turn our attention towards the application of networks within the field of biology. Systems biology is a field of study that seeks to model biological systems, emphasizing their interactions. Graphs and networks are widely

employed as effective representations of biological relationships. Various types of networks are used to capture specific aspects of biological interactions.

For example, gene co-expression networks are used to elucidate the relationships between genes exhibiting similar expression patterns. In such networks, nodes represent genes while edges represent co-expression relationships. These co-expression relationships can be quantified using various measures, depending on the methodology employed. For instance, the Weighted Gene Co-expression Network Analysis (WGCNA) often employs the Pearson correlation coefficient [56], while the Partial Correlation Information Theory (PCIT) method uses partial correlation [57]. These methods provide insights into the common involvement of genes in biological processes.

Gene regulatory networks (GRN) highlight the regulatory interactions between genes, shedding light on how genes influence each other's expression and function. In these networks, nodes represent genes and edges the activation or inhibition events [58].

Protein-protein interaction networks (PPI) depict the physical interactions between proteins. In these networks, nodes represent proteins and edges their physical interactions. This representation enables the comprehension of how proteins interact and form complexes to carry out their biological functions within cells. Notably, a well-established application is the guilt-by-association principle, which posits that highly interconnected proteins share functional properties and may be components of the same biochemical pathway [59, 60].

Metabolic networks provide insights into the interconnected metabolic reactions within cells. The aim is to study the production and transformation of small molecules through metabolism. Additionally, biological pathways networks are often used to represent and study sequences of biological reactions and interactions that contribute to specific functions or processes within a cell or organism [61]. In both types of networks, nodes typically represent metabolites or proteins, and edges symbolize biochemical and control reactions, facilitating a comprehensive understanding of cellular processes. Notably, well-known databases like KEGG [62] and Reactome [16] are prominent resources for accessing and exploring these networks.

There are other types of biological networks such as disease networks [63]. Using networks to represent biological interactions at large scale, systems biology offers a holistic and comprehensive view of biological systems. However, as highlighted in "Big biology: The 'omes puzzle" [64], "*Biology's central dogma is essentially a parts list. DNA codes for RNA, which codes for protein. That may give you three basic 'omes (genome, transcriptome and proteome), but life happens only because these parts work together.*" This is the rationale behind the emergence of a new type of network: integration of multi-omics data networks. These heterogeneous

networks aim to reconcile various -omic levels, with nodes representing different types of biological entities and edges representing the diverse relationships that link them.

Integration approaches Due to their effectiveness in representing -omics data at different scales, networks became a widely employed approach in the development of integration methods.

The most common applications is the search for active modules [38, 65]. A module can be defined as a specific set of molecules that collaborate to enable a common biological function [66]. An illustrative example is the Weighted Gene Co-expression Network Analysis (WGCNA) technique, which constructs gene co-expression networks based on transcriptomic data. As mentioned in the previous subsection, this method aims to identify modules of co-expression, wherein genes share similar expression patterns. Subsequently, these co-expression modules are linked to phenotypic and physiological traits of the studied individuals, enabling the identification of co-expressed genes modules associated with specific traits [56]. This example illustrates the integration of trancriptomic and phenotypic data but network-based integration seeks to combine multiple -omic levels.

Network-based integration requires the construction of multi-omics networks, either from experimental data or by combining different networks [67].There are two primary approaches to build the multi-omic network: relying only on experimental data from single-omics analysis, where interactions among various entities are derived directly from experimental results, or adopting a systems biology approach, which considers prior knowledge of the system. In the latter case, relevant entities and interactions for the studied system are gathered from (or mapped onto) specific databases. It is important to note that isn't a straightforward task and comes with challenges related to data correspondences, redundancy and interoperability. Resources for constructing this network include previously presented biological networks and publicly accessible databases. To interconnect different networks, entities can be linked based on known relationships that exist between them. For example, genes can be linked to the proteins they encode, allowing the connection of a co-expression network with a protein-protein interaction network. Alternatively, a metabolic network can be used to link proteins and metabolites. One can choose to extract only known relationships among the molecules of interest, implying the existence of direct connections, or consider the entire network and project our molecules of interest onto it.

The constructed network can be a heterogeneous network, i.e. an extension of simple networks where the nodes and edges can be of different types. Alternatively, it can involve a more

structured integration of different networks linked by particular edges across the different layers to form a multilayer network [68, 69]. Multilayer networks are highly suitable models for representing empirical systems, including transportation, social and biological networks. These networks are composed of multiple layers interconnected (or not) by inter-layer edges, with nodes shared or unique to different layers, interconnected by intra-layer edges. This diversity results in various types of multilayer networks as illustrated in Figure 2.3: layers of multiplex interconnected networks share the same set of nodes, while layers in general interconnected ones may not (inter-layer coupling); in interdependent interconnected network, layers exhibit different inter-layer relationships (inter-layer dynamics) [70]. More details are available in the github repository of MuxViz [1].

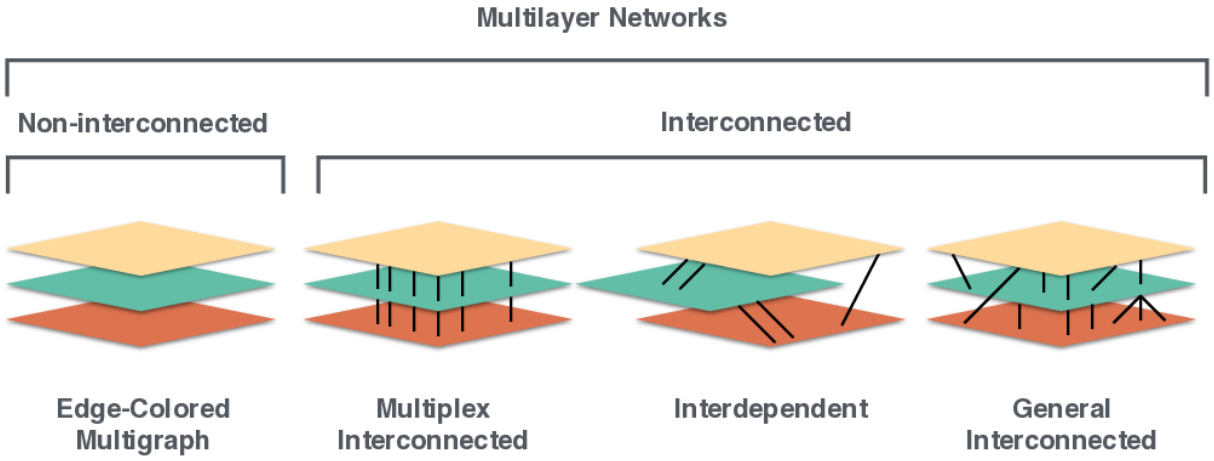

Figure 2.3 – Distinct classes of multilayer models. (Figure from Manlio De Domenico, "Multilayer Networks Illustrated" (2020) DOI:10.17605/OSF.IO/GY53K)

The obtained graph is often of considerable size and requires analysis by specific algorithms based on graph theory. Several methods can be employed to identify relevant connections or sub-networks. The advantage of this approach, compared to dimension reduction, lies in its incorporation of additional biological entities that potentially play crucial roles in the studied biological phenomenon. Another advantage of this approach is that the results are often easier to interpret.

Identifying relevant modules Once the biological network of interest has been identified or constructed, it can be analysed to uncover valuable insights into the organization and functioning of biological systems. Biological networks being modular because of the interplay of differ-

1. https://github.com/manlius/muxViz/blob/master/gui-old/theory/README.md

ent biological pathways in the whole network [66], methods have been developed to highlight these modules based on the network topology. Identifying a module consists in highlighting a subset (referred to as modules or clusters) of nodes that exhibit high interconnectivity within the network. This allows for unraveling the network's underlying structure.

Numerous methods have been developed to perform this analysis [65, 71–73], and their significance has been reviewed in [74]. Mitra et al. categorized these methods into three groups. Firstly, the significant-area-search methods guide the exploration through nodes and edges weighted with molecular activity scores. Modules are determined to optimize convergence of molecular activity scores inside modules. This approach gains its strength from a rich diversity strategies for scoring graph elements, coupled with various heuristics. Secondly, the diffusion-flow and network-propagation methods start from initial nodes and extend progressively their influence to neighboring nodes.Various implementations and formulations exist, including Random Walk with Restart (RWR) [75] and Heat Diffusion (HD) approaches [76]. Finally, the clustering-based methods involve grouping interactions in a network based on both their topology and the conditions in which they are active. For more comprehensive details and illustrative example, refer to [65].

These methods operate on single networks, and are used in single -omics layers or alternatively in aggregated multi-omic networks representing the interactions from the different sources. These methods treat all edges (interactions) equally, regardless of their specific molecular nature. This can be a limitation of these approaches. To address this gap, methods applicable across multiple networks and multiplex networks have emerged. Initially, MolTi and SimMod were developed for community detection within multiplex graphs [77, 78]. More recently, efforts have been made to improve these approaches on multilayer graphs based on the different redefinition of topological metrics on this type of network [69, 79]. Theses new methods include methods based on RWRs applied to multiplex networks [80, 81], or on multi-objective genetic algorithm such as in [82].

In the complex field of biological network analysis, obtaining meaningful insights goes far beyond statistical results. To make sense of this data, it's imperative to interpret it by contextualizing it within existing biological knowledge. However, crossing this crucial bridge requires efficient access to numerous available databases and the ability to query them precisely. This is where the semantic web and knowledge graphs stand as a pivotal technological framework, enabling the referencing of entities relative to one another across different sources. We will ex-

plore how this approach can be applied to biological networks, demonstrating how questions related to networks and graphs can benefit from this perspective, thus leading to a deeper and more precise understanding of biological interactions.

2.3 Knowledge integration and representation

2.3.1 Semantic Web, RDF Knowledge graphs and SPARQL

RDF (Resource Description Framework) is a formalism introduced by the W3C (*World Wide Web Consortium*) that provides a framework for representing and linking symbolic data in the form of a directed graph. It is *de facto* a standard for representing knowledge and information in a structured manner, enabling interoperability and data integration across different sources [83, 84].

In RDF, entities (or resources) and relations (or properties) are identified by unique URIs [2] (*Uniform Resource Identifier*) and data is organized into triples, which consist of *subject-predicate-object* statements.

— The *subject* represents the identifier of a resource (*i.e.* a node in the graph).

— The *predicate* denotes a specific property or relationship (*i.e.* a directed edge from the subject to the object in the graph).

— The *object* can either be a string value or the identifier of a resource (*i.e.* a literal value **or** a node). A string value can be post-fixed by a type to specify the data type of the value, for example while `"42"` should be considered as a string, `"42"^^xsd:integer` indicates that the value should be interpreted as an integer.

For example, the following code consists of five triples. The first triple states that the entity `uniprotkb:O96008` belongs to the category of proteins (indicated by the URI `up:Protein`; note that category descriptions are provided using the `rdf:type` instantiation relationship). The second and third triples respectively indicate that `uniprotkb:O96008` is marked as "reviewed" (which is a boolean attribute), and that it is linked to the entity `hgnc:18001` through the `rdfs:seeAlso` relationship. The fourth end fifth triples specify that `hgnc:18001` is an entity from the HGNC database and includes the comment "TOMM40" which corresponds to its common gene name. The corresponding graph is illustrated in Figure 2.4

2. URIs have been extended to IRIs (Internationalized Resource Identifiers) that support non-ASCII characters ; the principles remains unchanged.

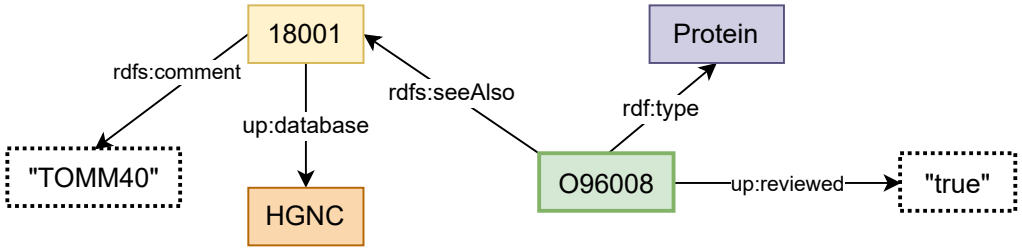

Figure 2.4 – Representation of information about the protein with UniProt identifier 'O96008' in RDF format. Entities are represented by solid-bordered rectangles, their attributes by dashed-bordered rectangles and are connected by properties represented by edges. The different triples form a directed graph.

```
1  PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
2  PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
3  PREFIX owl: <http://www.w3.org/2002/07/owl#>
4  PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
5
6  PREFIX up:<http://purl.uniprot.org/core/>
7  PREFIX udb: <http://purl.uniprot.org/database/>
8  PREFIX uniprot: <http://purl.uniprot.org/uniprot/>
9  PREFIX hgnc: <http://purl.uniprot.org/hgnc/>
10
11 uniprotkb:O96008 rdf:type up:Protein .
12 uniprotkb:O96008 up:reviewed "true"^^xsd:boolean .
13 uniprotkb:O96008 rdfs:seeAlso hgnc:18001 .
14 hgnc:18001 up:database udb:hgnc .
15 hgnc:18001 rdfs:comment "TOMM40" .
```

There are as many triples as needed to comprehensively describe a resource using literal values for specific attributes, RDF properties (such as `rdf:type`) to specify the type of the resource, and other properties to define relationships between resources. Consequently, RDF datasets are sets of triples. By using these triples, RDF can represent complex relationships and interconnected data in a flexible and extensible manner.

Additionally, the concept of "namespace" plays a crucial role. It serves as a mechanism to uniquely identify and organize resources and properties. URI are used to uniquely identify resources and these URI can be partitioned into namespace to gather related identifiers and prop-

erties (*e.g.* http://www.w3.org/1999/02/22-rdf-syntax-ns# for basic RDF resources and properties). This prevents naming conflicts when using several ontologies in the same RDF graph.

The graphs from several RDF dataset can be merged by simply performing the union of their triples. Moreover if these graphs use the same URI to identify a common resource, the graphs will e connected seamlessly.

The SPARQL language, which stands for SPARQL Protocol and RDF Query Language, is a query language used for searching, adding, modifying, and deleting data in RDF format. It is also based on the concept of triples and allows for the representation of variables (denoted by a preceding question mark).

The following query retrieves all molecular complexes that include the protein O96008 as one of their components. On a dataset containing complexes formed by this protein and others entities, it should return pairs of complex identifiers and its name. For example the pair (`reactome:R-HSA-1252240`, "TOMM40 Complex [mitochondrial outer membrane]") for the variables `?complexID` and `?complexName`.

```
PREFIX bp3: <http://www.biopax.org/release/biopax-level3.owl#>
PREFIX reactome: <http://www.reactome.org/biopax/84/48887#>

SELECT DISTINCT ?complexID ?complexName
WHERE {
?complexID bp3:component ?protein .
?protein bp3:entityReference/bp3:xref/bp3:id "O96008" .
?complexID bp3:displayName ?complexName .
}
```

A SPARQL endpoint is a web service that exposes an RDF dataset, enabling the execution of SPARQL queries and the retrieval of query results. The SPARQL engine returns combinations of variable values that satisfy the query constraints. SPARQL endpoint are valuable tools for accessing RDF data and conducting federated queries that span multiple distributed RDF repositories simultaneously. This capability is possible through the use of the SPARQL keyword `SERVICE`. This means that in a single query, it is possible to draw data from two or more repositories using shared features or identifiers.

RDF is therefore a well-suited formalism for solving data integration problems and for linking project data to public databases and knowledge bases [85–87]. In life sciences, most

databases and knowledge bases are available in RDF, supporting Life Science major place in the *Linked Open Data* [3].

Another advantage of RDF is its capability to integrate both data and knowledge by providing an unified framework based on triples to represent data and ontologies. Similarly, SPARQL allows the formulation of queries that operate on both aspects.

2.3.2 Graph databases: Neo4j, Cypher and Neosemantics

Neo4j [4] is classified as a Labeled Property Graph (LPG) database [89]. LPG databases are used to store data in a graph format, similar to RDF triple stores, but they employ a different data model compared to RDF. Unlike RDF, which uses triples and identifies resources using URIs, LPG databases rely on uniquely identifiable nodes and edges that can have associated properties. These properties are essentially key-value pairs, this concept is analog to RDF triples where the *object* is a literal value. This structural difference impacts the size and compactness of data, with LPG databases often being more compact than RDF graphs. The integration of data and knowledge is also more limited than with the Semantic Web, as complex relationships between node labels are not supported. Another significant difference between LPG and RDF is the absence of a concept like "namespace" in LPG. Indeed, LPG is primarily designed for storing and querying data within a single database system rather than for data exchange or integration across different sources. As a result, exploring data from multiple sources simultaneously in LPG databases can be less straightforward compared to RDF, which offers more robust mechanisms for data integration and linking across disparate sources. Nevertheless, the plugin NeoSemantics [5] enables to store RDF datasets in Neo4j and to export property graph from Neo4j as RDF graph. For instance the literal values of RDF that are not supported in Neo4j are transformed into key-value attributes of the subject node in Neo4j and vice versa.

There are several property graph query languages and Cypher is a well-established and widely used declarative one [90]. It was developed by the Neo4j group and is therefore highly optimized for exploring Neo4j graphs. Moreover graph databases support classical graph traversal and analyses algorithms (e.g. shortest paths, centrality, etc.) that are more difficult to implement with RDF/SPARQL.

RDF and SPARQL are well adapted to data integration and advanced reasoning based on

3. https://lod-cloud.net/ [88]
4. https://neo4j.com/
5. https://github.com/neo4j-labs/neosemantics

symbolic knowledge from ontologies, whereas graph databases are well adapted for complex analyses based on graph topology. It is interesting to note that my requirements encompass both aspects.

In this section, we have briefly introduced two approaches for representing knowledge in the form of graphs. On one hand, RDF graphs allow for highly structured knowledge organization through ontologies, which serve as models for data storage and enables interoperability between different data sources through federated queries and namespaces. On the other hand, graph databases like Neo4j and the query languages used for these databases allow for the expression of more complex queries than SPARQL for RDF. This is achieved by leveraging properties on nodes and edges, as well as graph theory. With the Neosemantics plugin, it is possible to combine these two technologies by transitioning from one model to another. This combination of technologies can provide a robust framework for handling and analyzing complex biological data, especially in the context of multi-omics research.

2.4 Case study: networks of entities associated with variability in feed efficiency of growing pigs

2.4.1 Context

Livestock production is significantly influenced by the cost of animal feed, which represents the largest portion of production costs. These feed costs are directly tied to the fluctuation prices of raw materials and subsequently have a direct impact on the income of farmers. In order to ensure the sustainability of farms and the industry as a whole, it is crucial to remain competitive and to adapt to international markets. However, as pig farming operations expand, they also face environmental challenges related to storage, treatment and management of effluents. Balancing economic viability with environmental sustainability has become the primary objective of the pork industry. To achieve this balance, various factors must be taken into account, and one of the most promising and widely considered strategies is to enhance feed efficiency of animals. Improving feed efficiency can optimize resource utilization, reduce feed costs, and minimize environmental impacts.

2.4.2 Definition

Feed efficiency in growing animals is defined as the relative ability of the animals to convert feed into weight (lean) gain while maintaining physiological functions. Improving feed efficiency directly contributes to reducing feed costs and minimizing the environmental impact of livestock production, particularly in terms of effluent discharge. By maximizing the utilization of nutrients from the feed, animals can achieve optimal growth while minimizing nutrient losses. Feed efficiency can be improved through various approaches, including genetic selection and research into high-performance feeds. However to further advance these strategies, it is crucial to gain a deeper understanding of the underlying biological processes that drive feed efficiency. Feed efficiency is indeed a complex phenotype influenced by numerous biological pathways, which makes measuring it a challenge in itself. Various metrics have been developed, with two of the most commonly used being Residual Feed Intake (RFI) [91] and Feed Conversion Ratio (FCR). FCR is the calculated by the ratio of feed intake to body weight (BW) gain during a test period. Consequently FCR is a measure based on production within livestock farming. A lower FCR suggests optimal growth rates and favorable body composition [92]. However as FCR is a ratio, it introduces some bias, an animal can exhibit efficiency by consuming less feed and still maintaining a lower growth rate. This observation highlights the multifaceted nature of feed efficiency where feed consumption and growth rate represent only one facet of this intricate phenotype. Residual feed intake (RFI) has been also proposed as a refined indicator of feed efficiency. Unlike FCR, RFI delves deeper into the complexity of the phenotype, it is calculated as the difference between observed feed intake and predicted intake from production and maintenance needs. This prediction requires a reference population, therefore, RFI is rather used as a genetic selection criteria. RFI is rather based on the metabolism efficiency than based on the daily gain and growth rate which allows RFI to reflect digestive and metabolic variabilities [91, 92]. It is important to note that there are other contributing factors beyond those captured by this two metrics. Moreover, alternative metrics may provide a more appropriate assessments based on the specific scientific question.

Feed efficiency stand as a key phenotype in sustainable agriculture, not only saving resources but also reducing waste and effluents into the environment. However, its complexity is profound, governed by a large number of biological pathways. Understanding this complexity is the first step towards its improvement. Recent studies showed that gene expression profiling in the whole blood is suitable to identify a few number of molecular candidate biomarkers for

FCR in growing pigs [93]. Likewise, metabolomic studies have shown that circulating concentrations of metabolites in the blood can be related to economically-important traits including feed efficiency [94].

CHAPTER 3

KNOWLEDGE DATABASES AND EXPERIMENTAL DATA USED IN THIS THESIS

In this chapter, we describe the materials used in this thesis, namely the UniProt and ChEBI knowledge bases and ontologies, the Reacome database, and experimental data from transcriptomics and metabolomics.

3.1 Databases

3.1.1 UniProtKB

The Universal Protein Resource Knowledgebase (UniProtKB) [1] [13] is a high-quality, freely accessible database containing protein sequences and functional annotations. It serves as a centralized and well-organized repository for protein-related information from various sources. UniProtKB consists of two main subsets:

— **UniProtKB/Swiss-Prot** contains manually annotated and expert-reviewed protein entries. Each entry is documented with information extracted from the literature. The data provided includes the protein's name, corresponding gene name, gene identifiers, organism details, functional annotations, roles in biological processes, enzymatic activities, post-translational modifications, domain structures, similar proteins, and various other properties.

— **UniProtKB/TrEMBL** (Translated EMBL Nucleotide Sequence Database) contains computationally predicted or inferred coding sequences. These sequences are sourced from databases such as Ensembl, RefSeq, and CCDS, etc.

Each protein within UniProtKB is assigned a unique entry name or identifier, making it easy to access and reference.

UniProt offers various methods for accessing and utilizing the UniProtKB including SPARQL

1. https://www.uniprot.org/

endpoint [2]. This endpoint allows users to query UniProtKB data using SPARQL queries, providing a powerful means to retrieve specific protein-related information and conduct customized analyses. Figure 3.1 presents an overview diagram illustrating the main concepts of the SPARQL endpoint. Starting from the accession number (UniProt ID) of a protein, users can access all the available information stored in the database navigating through the different properties and classes. Moreover, the diagram highlights the `rfs:seeAlso` property, which points to external databases. For example, these cross-references enable to retrieve UniProt IDs from a gene identifiers (e.g. HGNC), facilitating integration with external and/or experimental data.

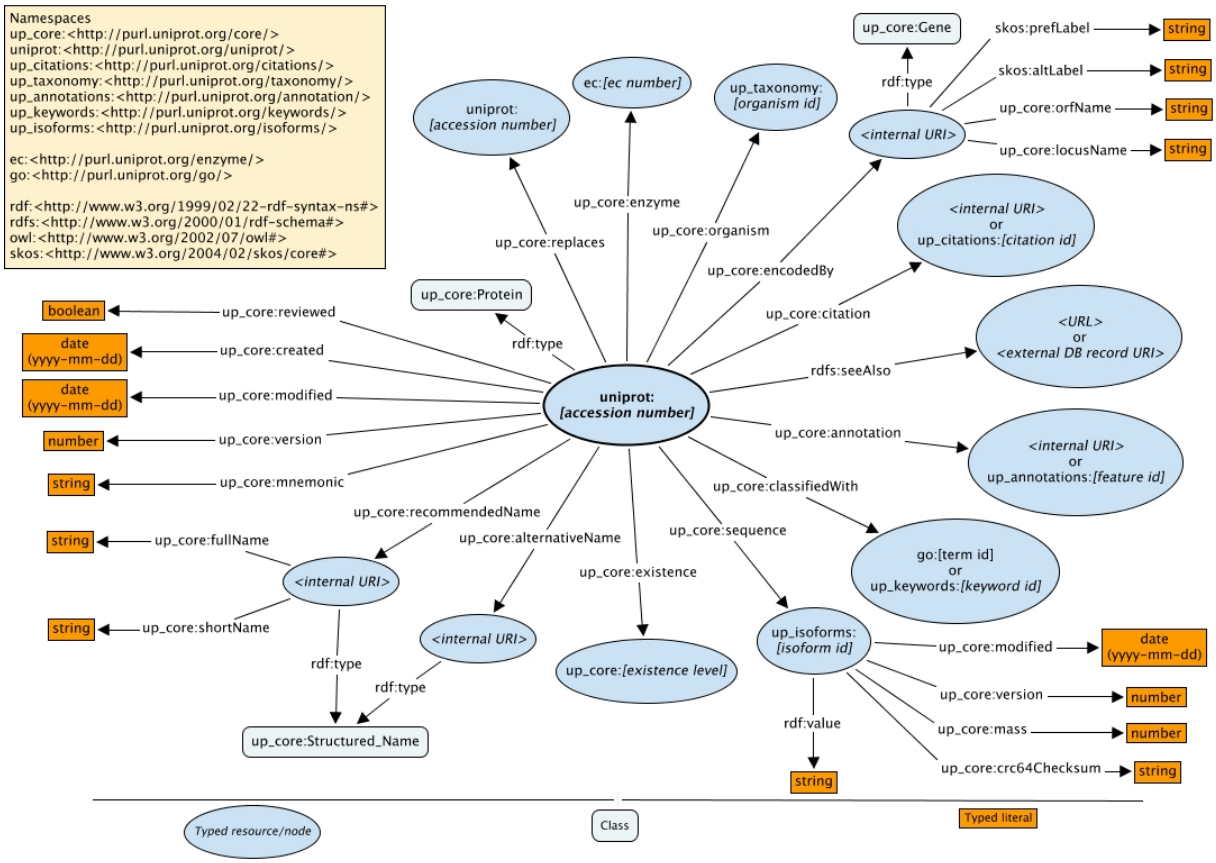

Figure 3.1 – Overview diagram on the main concepts of UniProt SPARQL endpoint. This diagram provides a visual representation of the interconnected data available through the UniProt SPARQL endpoint. (source: UniProt)

2. https://sparql.uniprot.org/

3.1.2 ChEBI

Chemical Entities of Biological Interest (ChEBI) [3] [18] is a freely accessible database and ontology dedicated to small chemical compounds. ChEBI is a part of the Open Biomedical Ontologies (OBO) project of the European Bioinformatics Institute (EBI). This ressource provides detailed information on small molecules, including details about their structures, properties and biological roles. ChEBI compiles and curates data from a wide range of sources, including the Integrated relational Enzyme database (IntEnz), the subset COMPOUD of the Kyoto Encyclopedia of Genes and Genomes database (KEGG), PDBeChem, ChEMBL (bioactive compounds) and numerous other repositories.

Each molecular entity within ChEBI is uniquely identifiable by a ChEBI identifier, which allows users to access to information. This information includes compound's common name, level of manual annotation, chemical formula, charge, synonyms, publications that reference it, a hierarchy tree linking it to other ChEBI entities, and cross-references to other resources.

The ChEBI ontology enables users to explore through related entities (subclasses, enantiomers, etc.) as well as navigate the hierarchy of biological roles using SPARQL queries.

3.1.3 Reactome

Reactome [4] [16] is an open access, manually curated and peer-reviewed database enriched with knowledge and mappings to third party ontologies. It focuses on biological pathways and processes. It serves as a centralized repository for knowledge on pathways extracted from literature. Reactome organizes these pathways based on the concept of reaction involving various entities. There entities are nucleic acids, proteins, complexes, vaccines, anti-cancer therapeutics and small molecules, all of which contribute to biological interactions that are grouped into pathways. Reactome provides data and pathways for various species, and each species is typically represented as a separate subdataset.

Reactome stores data in a relational database and provides multiple ways to access and use this information. For example, the Reactome website offers an interactive graphical representation of biological processes, making it easy to explore pathways visually. Futhermore, the data can be downloaded in various formats, ensuring accessibility for researchers who may wish to use them for their specific analyses. The available formats include Neo4j GraphDB, MySQL, BioPAX, SBML and PSI-MITAB files. Each of these formats has its own advantages and limits,

3. https://www.ebi.ac.uk/chebi/
4. https://reactome.org/

a subject that Strömbäck et al. studied extensively. They particularly emphasized the promising potential of the BioPAX format for finely describing biological pathways [95].

In this thesis, we mainly used exports from Reactome dataset in BIOPAX format, although in some occasions, we also used Neo4j. This decision aligns our goal of developing a generic method that is not dependent on the specific dataset. Additionally, Biological Pathway Exchange (BioPAX) format has been specifically designed to facilitate integration and interoperability among various sources of pathways data, promoting reproducibility in our work. It is important to highlight that major pathway databases, including KEGG and PathwayCommons, also provide data in the BioPAX format, which contributes to the genericity of our approach.

BioPAX is a well-established ontology to represent pathways at molecular and cellular levels as graphs using RDF and OWL technologies. As represented in Figure 3.2, this ontology is structured around the root class `Entity` which encompasses four key subclasses: `Pathway`, `Interaction`, `Gene` and `PhysicalEntity`. The `PhysicalEntity` class is the parent class for `Protein`, `DNA`, `RNA`, `SmallMolecule` and `Complex` classes. The `Interaction` class is used to describe biological relationships between entities. There are various subclasses of the `Interaction` class, each designed to represent different types of biological interactions. Some of these subclasses include: `Control`, `MolecularInteraction` and `BiochemicalReaction`. These subclasses may have further subclasses to provide finer-grained representation of specific types of interactions.

Entities are interconnected through various properties. For instance, the property `bp3:component` links `Complex` to its components, while `bp3:participant` (and its subproperties `left`, `right`, `controlled`, etc. depending on the interaction class) links `Interaction` to its participating entities. This structured representation facilitates the description of complex biological interactions and relationships within pathways.

BioPAX also includes utility classes like `EntityReference`, which serves as a means to group entities across different contexts and molecular states that share common physical properties. Each type of physical entities has a corresponding subclass of `EntityReference` (e.g. `ProteinReference`, `SmallMoleculeReference`). In practical terms, this means that for entities like proteins and small molecules, there is a dedicated node where all the non-changing aspects of the entity are stored. Figure 3.3 is an example of a `ProteinReference` (in red) that have a UniProt ID (O96008), and three proteins (in blue) that are linked to this node with the `entityReference` property. This linkage suggests that these three proteins share common attributes but may have variations, such as different cellular locations or molecular states.

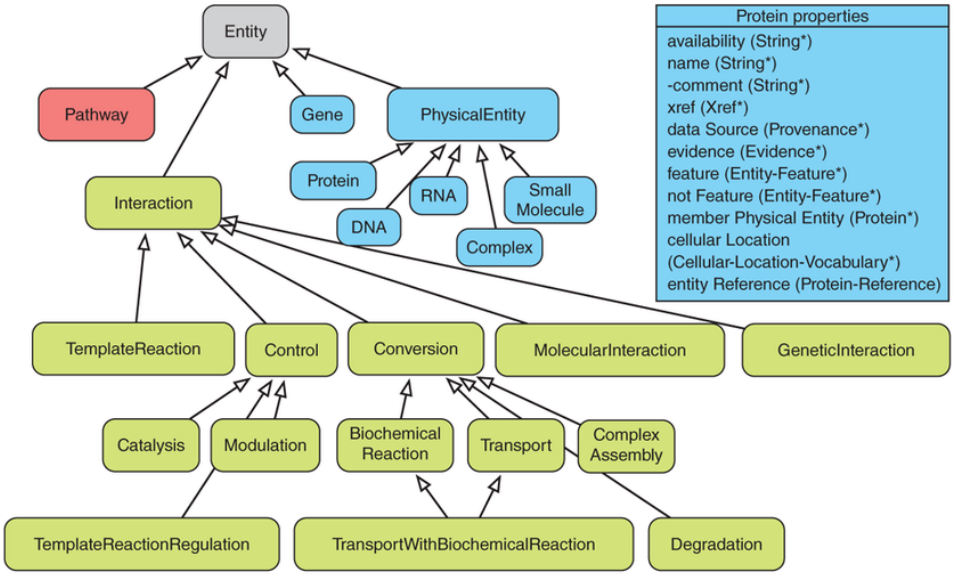

Figure 3.2 – An overview of the key classes in the BioPAX ontology, represented as boxes, and their inheritance relationships, indicated by arrows. There are three main types: Pathway (in red), Interaction (in green), and PhysicalEntity/Gene (in blue). To illustrate properties, the properties of the Protein class are displayed. (Figure extracted from Demir et al. [96])

BioPAX provides the capability to map entities to external resources through the use of the `XRef` utility class. There are several subclasses of "crossref" designed for different purposes. In the example provided in Figure 3.4, the `PublicationXref` class (in pink) is employed to reference the corresponding publication using PubMed ID. The `UnificationXref` class (in grey) is used to annotate biological entities, in this example it is used to reference the Reactome ID of the reaction. However, it can also be used to link to other resources like UniProt or ChEBI.

The BioPAX export of Reactome v.84 (2023-04) is composed of $493,858$ nodes and $2,885,960$ edges (triples), including $22,635$ `Interactions`, $31,332$ `Proteins`, $15,222$ `Complexes` and $5,083$ `SmallMolecules`.

3.2 Experimental data

In this thesis, we reused datasets acquired in the whole blood samples of 48 purebred French Large White pigs. These pigs were part of a divergent selection for Residual Feed Intake (RFI) and were fed two different diets (low fiber low fat and high fiber high fat), categorized into four equivalent groups (n = 12 by line and by diet). The age at slaughter ranges from 124 to 139 days, with an average of 132.7 days, a median of 132 days, and a standard deviation of 3.7 days.

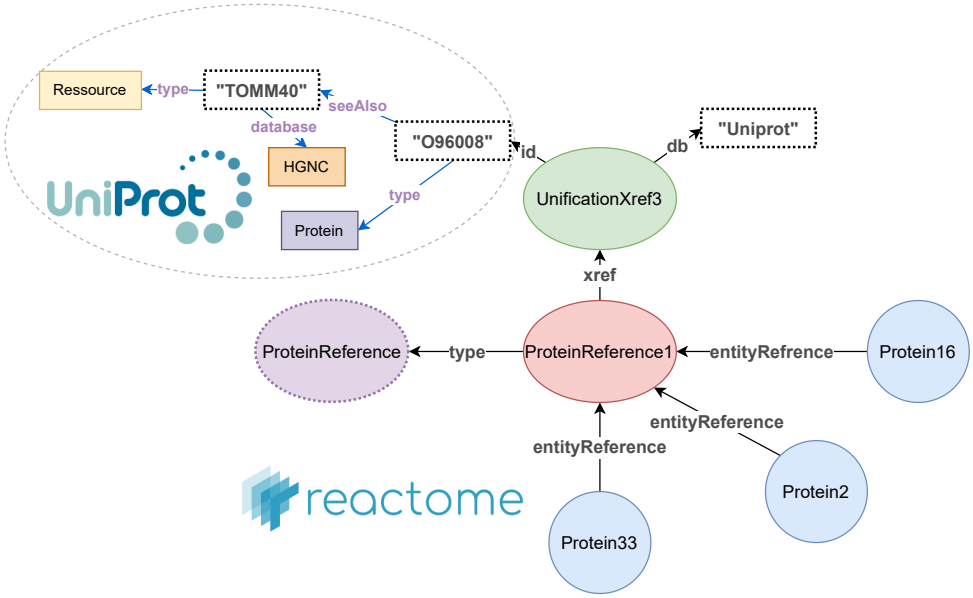

Figure 3.3 – Diagram illustrating the `EntityReference` concept of the BioPAX ontology. A `ProteinReference` instance is represented in red, with the UniProt identifier "O98008" obtained through cross-referencing with UniProt. Additionally, three `Protein` instances are connected to this protein reference, shown in blue, indicating that these three proteins are different instances of the protein "O96008".

The datasets encompass:

— **Transcriptomes in whole blood (RNA microarrays)** [97];

The gene expression in the whole blood is measured using 26,322 probes (approximately 20 base pair cDNA fragments), by using a custom porcine microarray (8x60K, GPL16524, Agilent Technologies France, Massy, France) containing 60,306 porcine probes. Regarding replicates, there are approximately 3 probes per selected gene. Data have been proceeded for filtration and were median-centered. During a prior analysis to detect outliers, one of the 48 pigs (number 41) was identified as aberrant because a large number of its probes have the same expression values. Therefore, pig number 41 has been removed from the dataset for all subsequent analyses.

— **Performance data**, including body weight (BW), average feed intake (ADFI), average daily gain (ADG), feed-conversion ratio (FCR) and various other measured phenotypic traits [98];

This dataset contains measurements for 15 distinct traits for each individual.

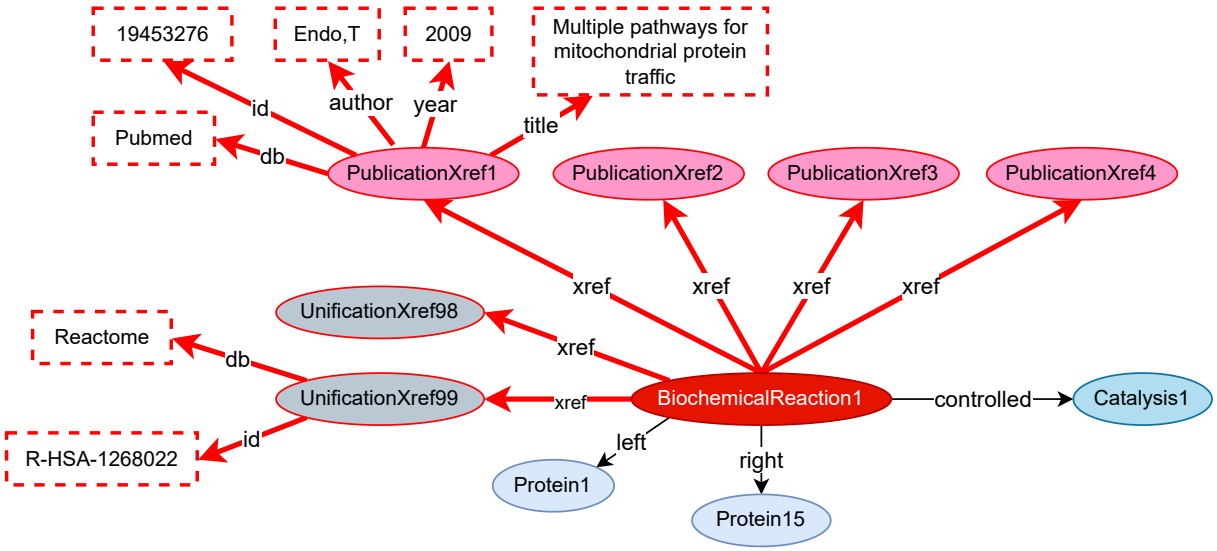

Figure 3.4 – Diagram illustrating the `Xref` concept of the BioPAX ontology using the example of a `BiochemicalReaction` (shown as a red node). All the properties and nodes related to this concept are highlighted in red. `BiochmicalReaction1` involves two proteins (depicted in blue) and controls an `Interaction` (depicted in turquoise). This `BiochemicalReaction` is linked to two type of `xRef` subclasses: `UnificationXref` (in grey) to establish cross-references with different databases by providing corresponding IDs, and `PublicationXref` (in pink) to link the publications related to this `BiochemicalReaction`.

— **Circulating concentrations of fatty acids (FA) in the plasma (gas chromatography)**; 14 concentrations were measured, including four groups of fatty acids categorized based on their functional properties: short and medium-chain saturated fatty acids, polyunsaturated fatty acids from the omega-6 family, polyunsaturated fatty acids from the omega-3 family, and very long-chain monounsaturated fatty acids.

— **Blood metabolome data (nuclear magnetic resonance (h1-NMR))** [97].
For each individual, 94 metabolites were identified, including amino acids and lipoproteins.

The original data have been analyzed in the referenced publications through differential analysis between line and/or diet.

CHAPTER 4

FIXING MOLECULAR COMPLEXES IN BIOPAX STANDARDS TO ENRICH INTERACTIONS AND DETECT REDUNDANCIES USING SEMANTIC WEB TECHNOLOGIES

This chapter has been published as an original research article to *Oxford Bioinformatics* (Juigné et al., 2023):

➡ **Juigné C**, Dameron O, Moreews F, Gondret F, Becker E. Fixing molecular complexes in BioPAX standards to enrich interactions and detect redundancies using semantic web technologies. Bioinformatics. 2023 May 4;39(5):btad257. doi: 10.1093/bioinformatics/btad257.

4.1 Abstract

Motivation: Molecular complexes play a major role in the regulation of biological pathways. The Biological Pathway Exchange format (`BioPAX`) facilitates the integration of data sources describing interactions some of which involving complexes. The `BioPAX` specification explicitly prevents complexes to have any component that is another complex (unless this component is a black-box complex whose composition is unknown). However, we observed that the well-curated Reactome pathway database contains such recursive complexes of complexes. We propose reproductible and semantically-rich SPARQL queries for identifying and fixing invalid complexes in `BioPAX` databases, and evaluate the consequences of fixing these non-conformities in the Reactome database.

Results: For the *Homo sapiens* version of Reactome, we identify 5,833 recursively defined complexes out of the 14,987 complexes (39%). This situation is not specific to the human dataset, as all tested species of Reactome exhibit between 30% (*Plasmodium falciparum*) and 40% (*Sus scrofa*, *Bos taurus*, *Canis familiaris*, *Gallus gallus*) of recursive complexes. As an ad-

ditional consequence, the procedure also allows the detection of complex redundancies. Overall, this method improves the conformity and the automated analysis of the graph by repairing the topology of the complexes in the graph. This will allow to apply further reasoning methods on better consistent data.

Availability: We provide a jupyter notebook detailing the analysis https://github.com/cjuigne/non_conformities_detection_biopax.

Supplementary information: Supplementary data are available at *Bioinformatics* online.

4.2 Introduction

4.2.1 Molecular complexes and biological interactions in system biology

Understanding how biological systems adapt to their environment requires to better capture, describe and model the interactions between their constitutive entities. With the accumulating knowledge on biological entities and their interactions, the need of a general framework to understand this information at a system level has led to a formal description of these entities and interactions. Within the context of biological pathways, several formats have been proposed such as SBML [99, 100], and `BioPAX` [96].

Biological systems typically involve an intricate network of interactions between numerous participants [101]. Among these participants, complexes are a major class of physical entities that results from the chemical assembly of several molecules (nucleic acids, proteins and other molecules) that bind each other at the same time and place, and form single multimolecular machines. Biologically, they play an important role in transcription, RNA splicing and polyadenylation machinery, protein export, transport [102, 103]. From the data analysis perspective, complexes cause some indirection between molecules and interactions, as molecules can participate directly to an interaction but also be a component of a complex that participates to an interaction. This introduces an additional node (the complex) between two entities that are no longer directly connected by a link in the cascade of events that triggers cell behaviour.

4.2.2 Description of complexes in BioPAX.

The Biological Pathway Exchange format [1] (BioPAX) is a well established formalism to represent biological pathways at the molecular and cellular levels, including interactions [96]. The BioPAX objective is to unambigously describe each component and each interaction. In the BioPAX ontology, the top four classes are Pathway, Interaction, Physical Entity and Gene. Interactions represent the biological relationships between two or more entities, including molecular interactions, controls and conversions. Physical entities encompass small molecules, proteins, DNA, RNA and complexes. Complexes are defined in BioPAX as "*physical entities whose structure is comprised of other physical entities bound to each other non-covalently, at least one of which is a macromolecule (e.g. protein, DNA, or RNA)*".

The BioPAX specification explicitly states "*complexes should not be defined recursively [...] i.e., a complex should not be a component of another complex. [...] Exceptions are black-box complexes (i.e., complexes in which the component property is empty), which may be used as components of other complexes because their constituent parts are unknown*". This specification about the flat representation of complexes was introduced when moving from the BioPAX1.0 to the BioPAX2.0 standard (between April and December 2005), and the rationale given for the introduction of this constraint was that the use of a tree structure could be interpreted by some users as an order in macromolecular assembly: *"The reason for keeping complexes flat is to signify that there is no information stored in the way complexes are nested, such as assembly order. Otherwise, the complex assembly order may be implicitly encoded and interpreted by some users, while others created hierarchical complexes randomly, which could lead to data loss."*

4.2.3 The Reactome use-case.

BioPAX is based on Semantic Web technologies, with RDF facilitating integration, SPARQL facilitating querying and OWL facilitating knowledge-based reasoning. All the major pathway databases are available in BioPAX. Among them, Reactome [2] is a free, open-source, curated and peer-reviewed pathway database [16]. It is widely used in genome analysis, modeling, systems biology, clinical research and education, and biological pathways can be explored to shed ligth on interconnected proteins [104]. Despite the BioPAX specifications, we noticed the presence of recursive complexes, i.e. complexes composed of other complexes that are not black-

1. http://www.biopax.org/release/biopax-level3-documentation.pdf
2. https://reactome.org/

box complexes. Importantly, these non-conform complexes were not detected by the `BioPAX` validator [3] [105]. As this pattern eludes validation, their presence in Reactome and possibly in other databases, may preclude further analyses aiming to provide robust information about mechanisms and phenotypes.

4.2.4 Motivations and results

We hypothetized that identifying and correcting recursive complexes would be valuable to enrich the biological database into interactions. This may allow a better analysis of the participants, direction and stoichiometry of the biological interactions, either when browsing the data, or when data are processed automatically. We believe that these non-conformities in the description of the complexes are an important obstacle to the development of analysis methods that would work directly from the `BioPAX` format. Using SPARQL queries, we show that the well curated Reactome database includes a large fraction of recursively defined complexes, *i.e.* whose components contain at least one complex (that is not a black box one). We showed that these non-conform complexes averaged one-third of the total number of complexes. Using other SPARQL queries, we corrected these recursive decompositions of complexes. Then, we showed that these corrections led to the detection of implicitly redundant complexes, whose redundancy was previously hidden by the different recursive definitions of complexes.

4.3 Approach

4.3.1 Definition of invalid recursive complexes

Recursive complexes are complexes whose component contains at least one complex. These recursive complexes are invalid if the inner complex is not a black-box one, *i.e.* the inner complex itself contains at least one component. The different categories of (in)valid complexes in `BioPAX` are illustrated in Figure 4.1-A.

4.3.2 Invalid recursive complexes in interactions

Recursive complexes can cause false negatives when identifying the interactions in which a physical entity can participate. For example, if A, B and C are physical entities, A can directly

3. https://biopax.baderlab.org/

participate in several interactions, but also indirectly when associated to B as a complex, or to B and C as another complex. If (A, B) and (A, B, C) are valid complexes, the two situations can be correctly processed by identifying the interactions matching the criterion “having a participant that is a complex composed of A". However, if (A, B, C) complex is composed of the complex (A, B) and of C, all the interactions in which (A, B, C) participates would fail to meet the aforementioned criterion, because their participants are not “a complex composed of A" but “a complex composed of a complex composed of A". We will see that in practice, such nested composition can occur over multiple levels. The approach we used to identify and fix these invalid recursive complexes, while respecting the stoichiometry where available, is illustrated in Figure 4.1-B.

4.3.3 Redundancies

As an additional consequence, fixing invalid recursive complexes can also result in the identification of redundancies between complexes, previously hidden by different recursive decompositions of complexes. For example, as illustrated in Figure 4.1-C, complexes $(A, (B, C))$, $((A, C), B)$ and $((A, B), C)$ could be distinct (invalid) complexes with their own identifier and at first glance, having different participants. However, they could be all fixed as a single complex having the same components A, B and C.

4.4 Methods

We developed semantically-rich SPARQL queries for identifying and fixing invalid recursive complexes, and detecting the resulting redundancies. We applied this method on the Reactome pathways database as a use-case study (version 81 (2022-06-13)). For the sake of reproducibility, we provide a jupyter notebook detailing the analysis [4].

4.4.1 Identifying invalid complexes

We first analyzed all the configurations of `BioPAX` complexes according to the nature of their components (4.1-A). The SPARQL query presented in figure 4.2 allows to identify the invalid recursive complexes. Other SPARQL queries to identify non-recursive complexes and valid recursive complexes, are presented in the Jupyter notebook.

4. https://github.com/cjuigne/non_conformities_detection_biopax

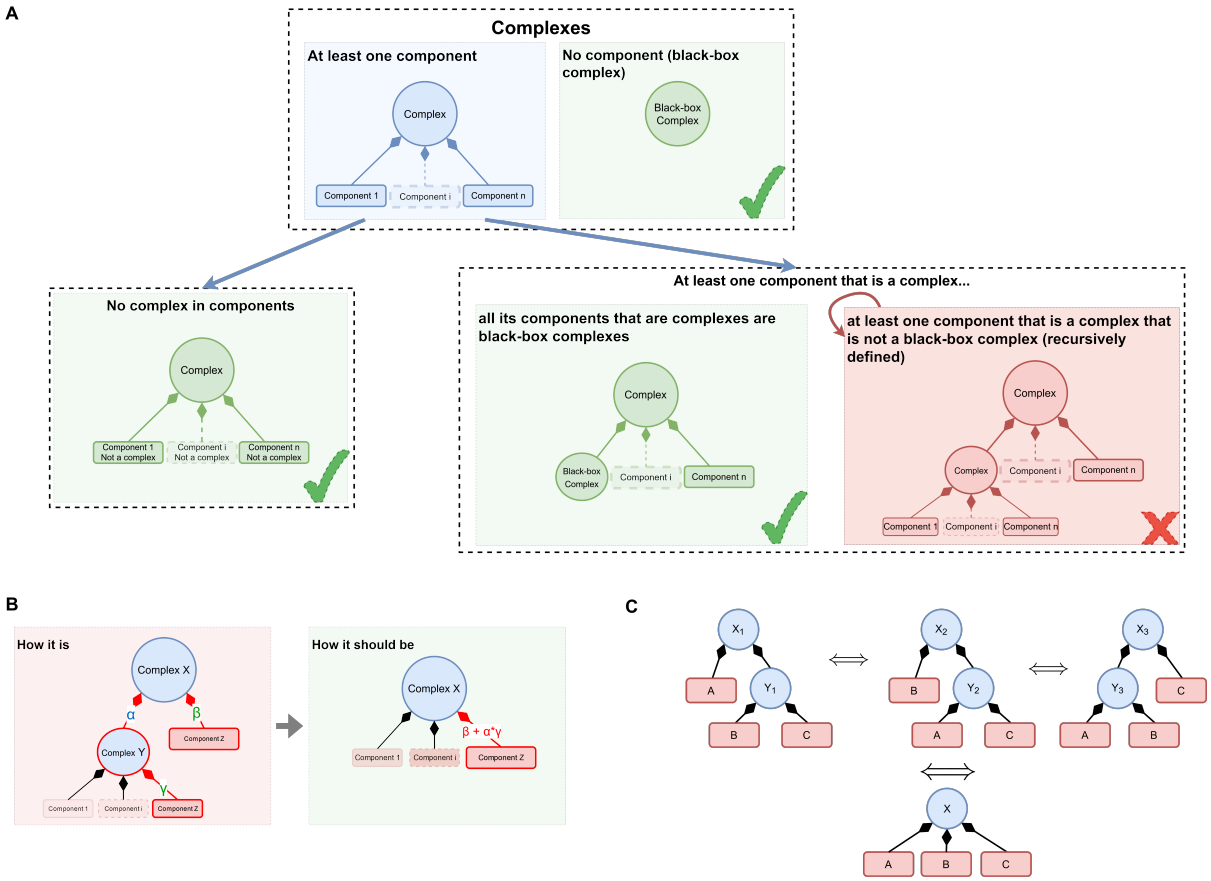

Figure 4.1 – Illustration of the conformity of complexes with respect to the BioPAX specifications: identification, correction and redundancies. (A) Validity and invalidity of the categories of BioPAX complexes. Complexes are represented by circles. Composition is represented by a diamond head arrow from the component to its complex. A valid complex can have components that are themselves complexes only if these complexes are all black-box complexes, i.e., they do not have any components. Note that an invalid complex can itself be a component of another complex, which therefore becomes invalid as well. (B) Fixing an invalid recursive complex consists in collapsing as direct components all its direct and indirect components that are leaves in the composition tree of the complex, and then, computing the correct stoichiometric coefficient values with equation 4.1. (C) Example of invalid recursive complexes leading to redundancy in the database. Fixing them made possible the detection of redundancy.

4.4.2 Fixing the invalid complexes

Fixing an invalid recursive complex can be decomposed with a four steps methodology: (1) collapsing as direct components all its direct or indirect atomic components (i.e., those that are not complexes or black-box complexes); they correspond to the leaves in the tree of components, (2) deleting all the other components, (3) setting the correct values for the stoichiometric

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX bp3: <http://www.biopax.org/release/biopax-level3.owl#>

SELECT DISTINCT ?invalidComplex
WHERE {
 ?invalidComplex rdf:type bp3:Complex .
 ?invalidComplex bp3:component ?complexComponent .
 ?complexComponent rdf:type bp3:Complex .
 ?complexComponent bp3:component ?componentOfComplexComponent .
}
```

Figure 4.2 – SPARQL query to identify invalid recursive complexes in BioPAX, i.e., the complexes composed of at least another complex that has components.

coefficients, and (4) preserving all other attributes of the complex. The whole procedure consisting of python scripts and SPARQL queries is available in the Jupyter notebook.

To collapse the direct components of the complex in steps (1) and (2), the original BioPAX relation component was replaced by a component relation between the root complex and its leaves in the tree of components. We kept all the other relations of the complex in step (4).

To compute the stoichiometric coefficients in step (3), we traced the stoichiometric coefficients from each leaf up to the root complex. We also considered the fact that a physical entity can be a component of several parts of the recursive complex. Tracing stoichiometry is illustrated by Figure 4.1-B where complex Y was composed of γ Z, and X was itself composed of α Y. This resulted in $\alpha \times \gamma$ Z in X (*via* Y). If, in addition to being a component of Y, Z was also a direct component of X with β as stoichiometric coefficient value, this resulted in $\alpha \times \gamma + \beta$ occurrences of Z in X.

We noted $S_y(z)$ the global stoichiometric coefficient value of Z at Y, i.e. the number of occurrences of Z in Y, and $C(y)$ the set of the direct components of Y. Formula 4.1 recursively computes the stoichiometric coefficient value of any physical entity Z.

$$\begin{cases} S_z(z) = 1 & \\ S_y(z) = 0 & \text{if } (y \neq z) \wedge (C(y) = \emptyset) \\ S_y(z) = \sum_{p \in C(y)} S_y(p) \times S_p(z) & \text{otherwise} \end{cases} \tag{4.1}$$

4.4.3 Identifying redundant complexes

We considered as redundant the complexes that have exactly the same components with the same stoichiometric values and the same cellular location. Figure 4.1-C illustrates how invalid recursive complexes can be the cause of redundancy due to the order by which the components are nested in the complexes. Fixing the invalid complexes made possible the detection of redundancy. For that, we developed a SPARQL query that identifies the pairs of complexes that have the same components and properties but different identifiers. This query is also available in the Jupyter notebook.

4.5 Results

4.5.1 Invalid complexes represent a significant part of complexes in the `BioPAX` description of Reactome

The Human subset of Reactome v81 is composed of a total of $14,987$ complexes. Among them, we identified 862 black-box complexes (*i.e.* complexes without any components). Among the remaining $14,125$ complexes with at least one component, $8,292$ complexes have no component that is itself a complex with components. Together with the 862 black-box complexes, they represent the $9,154$ valid complexes.
On the opposite, we identified $5,833$ complexes that have at least one component that is a complex with components. Altogether, these $5,833$ invalid recursive complexes represent 39% of the 14,987 complexes in Reactome. None of them have been detected by the BioPAX-validator tool [105].
These invalid recursive complex participate to $7,149$ out of the $22,237$ interactions identified in Reactome (i.e., 32%).

4.5.2 Fixing invalid complexes increases the average number of components participating to complexes in Reactome

All the $5,833$ invalid complexes were fixed thanks to a python script and SPARQL queries (4.4.2). After fixing recursive complex description, all components of the complexes are described by a flat representation in accordance with the `BioPAX` specification.

As expected, with this flat representation, the number of direct components implicated in a complex increases (paired t-test, $p < 0.0001$). Indeed, in the initial Reactome dataset, the average number of direct components in a complex is 2.2 ($\sigma = 2.6$) and the complexes with the largest number of components are `R-HSA-5626171` and `R-HSA-72069`, each having a maximum of 65 components. After fixing the invalid complexes, the average number of direct components in a complex is 4.3 ($\sigma = 8.7$) and the largest complex is `R-HSA-156656` with 151 components. The most drastic changes concerns complexes `R-HSA-927767` and `R-HSA-927890` which both move from 3 to 103 direct components. Figure 4.3 illustrates the number of direct components identified before and after fixing the invalid complexes. The distribution of the gain in the number of direct components is available in supplementary files - Figure 1.

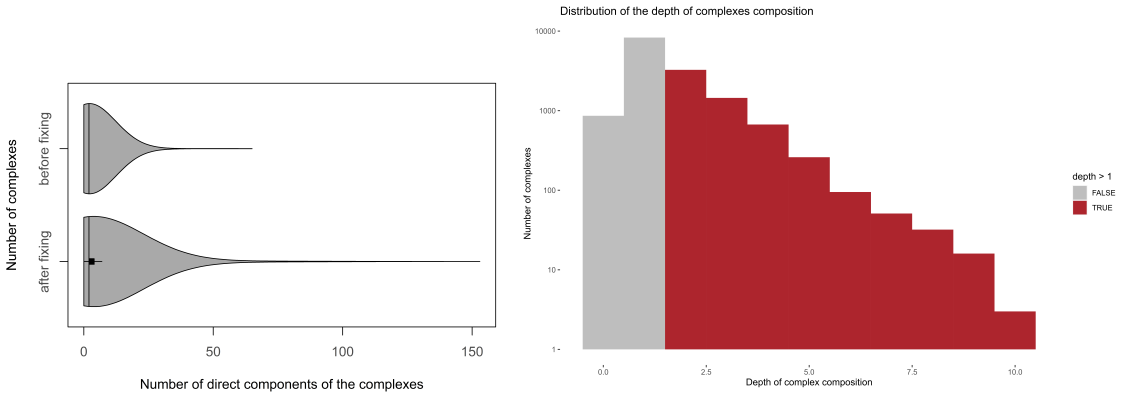

Figure 4.3 – Comparison of topological features before and after fixing invalid complexes. (A) Number of direct components of the complexes in the Human dataset of Reactome v79 before (top) and after (bottom) fixing invalid recursive complexes. Fixing invalid complexes clearly increased the number of components counted in many complexes. (B) Distribution of the depth of recursive complex definitions in the initial Human dataset of Reactome v81 (logarithmic scale).

4.5.3 Fixing invalid complexes reduces the path length from a complex to each of its components

We studied the composition depth of the complexes, before and after fixing the invalid ones. For a given complex, its composition depth was measured as the maximal length path between the root complex to its leaves. As illustrated in figure 4.3-A, the tree-like definition of invalid complexes artificially increases the depth of the complex composition.

In the figure 4.3-B, the depth of complex composition is represented before the fixing proce-

dure. Valid complexes have a depth of either one, or zero when they are black-box complexes. The red part of the figure represents invalid complexes having a depth greater than one. The maximum depth is 10, which leads to an artificial extension of the path length from the root complex to the majority of its leaves (example of R-HSA-68466 is given in supplementary file - Figure 2.

As expected, the correction of invalid complexes repairs the topology by reducing the path length between a complex and its components to a maximal length of 1. The depth of all complexes is 0 (black-box complexes) or 1 (complexes with components).

4.5.4 Fixing invalid complexes improves the detection of redundant complexes

Redundancies are detected between entity pairs but can also occur between more than two entities, as illustrated in figure 4.1-C with 3 equivalent complexes . In this example, 3 pairs of redundancies (X_1, X_2), (X_2, X_3), (X_1, X_3) are thus detected, corresponding to the size of a maximal clique with 3 vertices.

Among the $14,987$ complexes from the original Reactome database, the SPARQL query identifies 137 pairs of redundant complexes involving 249 distinct complexes. They constitute 121 maximal cliques of equivalent complexes (clique size ranging from 2 to 6 complexes), corresponding to 128 complexes in excess. In other words, we highlighted 121 groups of 2 to 6 redundant complexes identified by searching for pairs of redundant complexes. These redundancies are explicit, as they are not masked by a tree-like definition of complex components.

After fixing the invalid complexes, we identified 217 pairs of redundant complexes involving 347 distinct complexes. They constitute 164 maximal cliques of equivalent complexes (clique size ranging from 2 to 6), corresponding to 183 complexes in excess. The fixing procedure, by replacing the tree-like definition of complexes by a flat description of complexes, thus allows to detect more redundancies. These redundancies are implicit, as they are masked by a tree-like definition of complex components. They become explicit with the flat description of complex components. Figure 4.4 illustrates how fixing structurally different complexes reveals this redundancy.

Reactome contains cross-references to ComplexPortal [106] to annotate complexes. We sought to verify the possibility of identifying complex redundancies in Reactome without our SPARQL query-based procedure, by simply and unambiguously identifying complexes via their identifiers in a specialized database dedicated to complexes such as ComplexPortal, even if

ComplexPortal incompletely supports stoichiometry, which may be a severe limitation. Because of the very modest size of the ComplexPortal Database (1429 complexes for Homo sapiens), ComplexPortal IDs do not allow to detect any redundancies identified with the SPARQL query : among the 347 complexes that constitute 217 pairs of redundant complexes identified, only 3 out of 347 have a ComplexPortal ID.

This study is available in a Jupyter notebook on the GitHub repository. The supplementary table 1 lists all symetric `hasSameCompositionAs` relations between cliques of redundant complexes. The corresponding `.ttl` files are available on the GitHub repository [5].

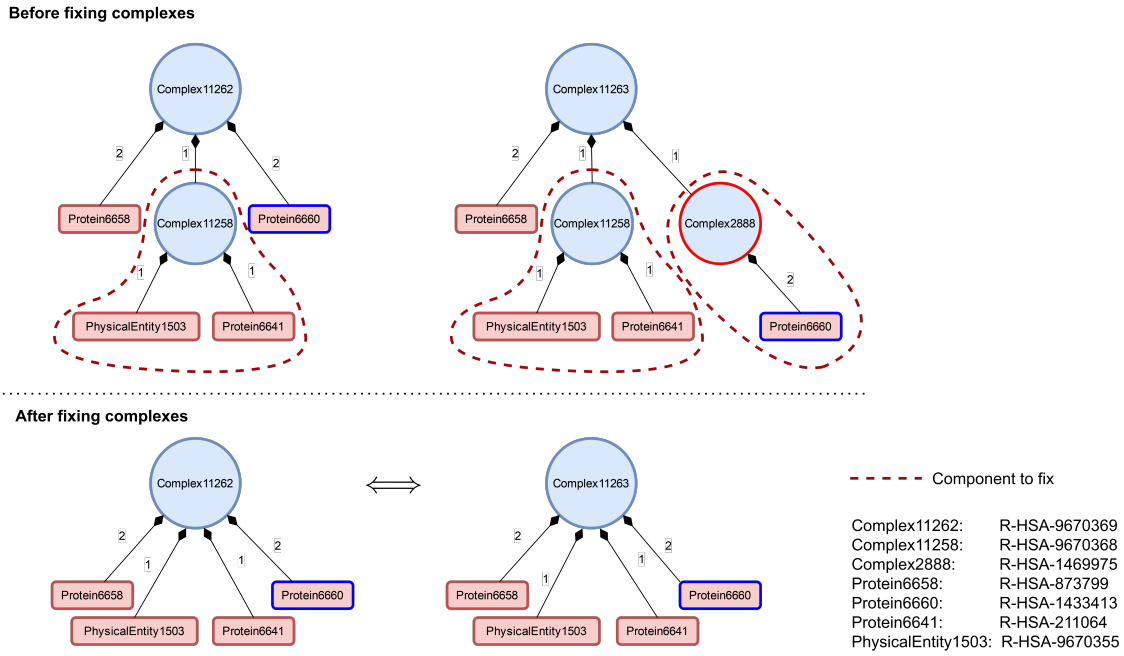

Figure 4.4 – Original invalid compositions of `Complex11262` (`R-HSA-9670369`) (top left) and `Complex11263` (`R-HSA-9670372`) (top right) in Reactome. Some components are complexes that should be replaced by a flat definition of components (indicated with dashed red lines). The fixed versions (bottom left and right, respectively) have a greater number of direct components than the original. Both fixed versions have the same components with the same stoichiometric coefficients, which reveals their redundancy. The structure difference between the original versions is highlighted in red: `Complex2888` (`R-HSA-9670372`) is composed of an intermediate dimer of the p-STATs protein `Protein6660` whereas `Complex11262` (`R-HSA-9670369`) is directly composed of p-STATs `Protein6660` with a stoichiometric coefficient of 2. A more detailed figure is available in supplementary files - Figure 3.

5. https://github.com/cjuigne/non_conformities_detection_biopax

Organism	Total number of complexes	(4.1) Invalid complexes	(4.2) Average number of direct components		(4.3) Redundancy	
			before fixing	after fixing	before fixing	after fixing
homo sapiens	14987	5833 (39%)	2.2 (std 2.6) max 65	4.3 (std 8.7) max 151	137 (121 cliques)	217 (164 cliques)
mus musculus	10707	4235 (39%)	2.3 (std 2.9) max 65	4.5 (std 9.0) max 151	2 (2 cliques)	16 (16 cliques)
sus scrofa	9022	3638 (40%)	2.3 (std 2.9) max 65	4,7 (std 9,4) max 151	0 (0 clique)	12 (12 cliques)
bos taurus	9412	3773 (40%)	2.3 (std 2.9) max 65	4.6 (std 9.2) max 147	0 (0 clique)	12 (12 cliques)
saccharomyces cerevisiae	1662	517 (31%)	2.5 (std 3.3) max 50	6.0 (std 12.4) max 106	1 (1 clique)	5 (5 cliques)
caernorhabditis elegans	4350	1560 (36%)	2.4 (std 3.3) max 64	5.0 (std 10.4) max 149	0 (0 clique)	8 (8 cliques)
canis familiaris	8945	3601 (40%)	2.3 (std 3.0) max 65	4.7 (std 9.4) max 151	0 (0 clique)	12 (12 cliques)
danio rerio	8618	3391 (39%)	2.3 (std 3.0) max 65	4.7 (std 9.5) max 150	0 (0 clique)	11 (11 cliques)
dictyostelium discoideum	2366	792 (33%)	2.4 (std 2.8) max 50	5.3 (std 9.8) max 103	0 (0 clique)	3 (3 cliques)
drosophilia melanogaster	5361	1955 (36%)	2.4 (std 3.0) max 64	4.8 (std 9.7) max 149	2 (2 cliques)	9 (9 cliques)
gallus gallus	8046	3244 (40%)	2.3 (std 2.8) max 65	4.7 (std 9.4) max 149	2 (2 cliques)	13 (13 cliques)
plasmodium falciparum	875	264 (30%)	2.4 (std 3.6) max 50	5.4 (std 11.9) max 103	0 (0 clique)	2 (2 cliques)
rattus norvegicus	9645	3780 (39%)	2.3 (std 2.9) max 65	4.5 (std 9.2) max 151	4 (4 cliques)	15 (15 cliques)

Table 4.1 – For each organism in the Reactome database, we counted the total number of complexes, then we identified the number of invalid complexes and evaluated the average number of direct components and the number of redundant complexes (pairs and cliques), before and after fixing the invalid complexes.

4.5.5 Application to non-Human organisms in the Reactome database

The complete procedure was then applied to all other organisms available in Reactome, to determine whether the large fraction of non-conform complexes detected in 4.5.1 (39%) is specific or not to the well-explored Human dataset. The results are presented in table 4.1. This shows that the large fraction of invalid complexes is not restricted to the Human dataset but also accounts for all of the 13 tested species. Datasets from all species exhibit between 30% (*Plasmodium falciparum*) and 40% (*Sus scrofa*, *Bos taurus*, *Canis familiaris*, *Gallus gallus*) of recursively defined invalid complexes. For all these species, repairing the topology of the complexes in the same way as for *Homo sapiens* significantly increases the average number of direct components of the complexes. With the flat representation of complexes, we also observed complex redundancies in complexes for each organism, although not in the same proportion as in the Human dataset (from 2 cliques for *Plasmodium falciparum* to 16 cliques for *Mus musculus*).

4.6 Discussion and perspectives

In this study, we show that non-conform recursive complexes affect a large proportion of Reactome database exported in the `BioPAX` format. Indeed, they constitute 30 to 40% of the complexes for all organsims, and participate to about one third of the interactions. The fact that this phenomenon occurs in all Reactome organisms may be explained in part by the fact that some interactions and complexes in non-human species are inferred if a large fraction (at

least 75%) of the proteins involved in the interactions or complexes have an ortholog for the species considered in PANTHER. Thus, taking into account corrections for Homo Sapiens prior to PANTHER inference process will probably result in a diminution of the phenomenon for other species. Due to this pervasiveness, any solution to overcome or repair recursive complexes would be valuable for automated graph analysis whatever the biological questions to be adressed.

In addition to being widely present in the datasets, we also show that invalid complexes composition reaches up to 10 levels in the tree of components of the complex. In these situations, navigating in the BioPAX file from some components of the complete complex to the complete complex is both painstaking and detrimental to computational performance.

Fixing recursive complexes consists in adding all the indirect components of a complex as new direct components. This leads to two main outcomes. First, it helps to reduce the path length from a complex to its components, since the maximal path length with the conform topology is now 1 (which solves the navigation limitation identified previously). Second, this increases the average number of components participating to complexes in Reactome. In the human dataset, this correction doubled the average number of direct components of a complex, from 2.2 to 4.3 components. This result is not specific to the human dataset: for all other organisms, the number of direct components has doubled.

As a side effect, the procedure reveals redundancies between complexes. These redundancies would have been difficult to identify without the procedure, because they are masked by the recursive definition of complexes. This is particulary relevant for the Human dataset, since the redundant complexes increase from 137 to 217 redundant pairs (+58%) before and after the fixing procedure. For the identification of redundant complexes, the verification that the cellular location is the same is crucial. Indeed, without considering cellular location, 243 pairs of redundant complexes can be identified whereas taking into account cellular location reduces this number to 217 pairs of redundant complexes. This difference is caused by the fact that physical entities can exist several times in BioPAX as long as they refer to physical entities in different states (including modifications, cellular location, etc.).

The BioPAX ontology, as defined in [96], is a very powerful format to represent and unify all the subtile levels of interactions occuring in biological pathways. It exploits the ontology formalism to conciliate genericity (with the high level entity classes Physical Entity and Interactions) and a precise description of the processes (with low level classes and various properties). However, this is also a quite complicated format description, and induces a computational complexity when reasoning on the data structured in this format. Strömbäck et al.

anticipated this, stating that: “This makes it possible to benefit from reasoning and conclusions based on the semantics given by OWL and the ontology, but the cost is a higher computational complexity for reasoning and integration of data." [95]. We have showed that more recent computational advances such as SPARQL, which is designed to handle ontologies, are better adapted to this task and can be also scale up.

Taking advantage of SPARQL queries, the procedures developed herein to (i) fix recursively defined complexes and (ii) identify redundant complexes, allow to correct the non-compliance with `BioPAX` specifications. As these procedures are applied at the level of the `BioPAX` files, they can be applied to all other major `BioPAX` pathway databases, including Kegg [62], MetaCYC [107], PathwayCommons [108], and WikiPathways [109], to assess the importance of invalid recursive complexes in these databases. Our strategy to modify directly the `BioPAX` files also ensures that further analyses of biological networks can be processed without any needs to modify any standard queries or scripts based on `BioPAX` libraries such as Paxtools [110] or PyBioPAX [111].

CHAPTER 5

SMALL NETWORKS OF EXPRESSED GENES IN THE WHOLE BLOOD AND RELATIONSHIPS TO PROFILES IN CIRCULATING METABOLITES PROVIDE INSIGHTS IN INTER-INDIVIDUAL VARIABILITY OF FEED EFFICIENCY IN GROWING PIGS

This chapter has been published as an original research article to *BMC Genomics* (Juigné et al., 2023):

➡ **Juigné C**, Becker E, Gondret F. Small networks of expressed genes in the whole blood and relationships to profiles in circulating metabolites provide insights in inter-individual variability of feed efficiency in growing pigs. *BMC Genomics* **24**, 647 (2023). https://doi.org/10.1186/s12864-023-09751-1.

5.1 Abstract

Background Feed efficiency is a research priority to support a sustainable meat production. It is recognized as a complex trait that integrates multiple biological pathways orchestrated in and by various tissues. This study aims to determine networks between biological entities to explain inter-individual variation of feed efficiency in growing pigs.

Results The feed conversion ratio (FCR), a measure of feed efficiency, and its two component traits, average daily gain and average daily feed intake, were obtained from 47 growing pigs from a divergent selection for residual feed intake and fed high-starch or high-fat high-fiber diets during 58 days. Datasets of transcriptomics (60 k porcine microarray) in the whole blood and metabolomics (1H-NMR analysis and target gas chromatography) in plasma were available for

all pigs at the end of the trial. A weighted gene co-expression network was built from the transcriptomics dataset, resulting in 33 modules of co-expressed molecular probes. The eigengenes of eight of these modules were significantly ($P \leq 0.05$) or tended to be ($0.05 < P \leq 0.10$) correlated to FCR. Great homogeneity in the enriched biological pathways was observed in these modules, suggesting co-expressed and co-regulated constitutive genes. They were mainly enriched in genes participating to immune and defense-related processes, and to a lesser extent, to translation, cell development or learning. They were also generally associated with growth rate and percentage of lean mass. In the whole network, only one module composed of genes participating to the response to substances, was significantly associated with daily feed intake and body adiposity. The plasma profiles in circulating metabolites and in fatty acids were summarized by weighted linear combinations using a dimensionality reduction method. Close association was thus found between a module composed of co-expressed genes participating to T cell receptor signaling and cell development process in the whole blood and related to FCR, and the circulating concentrations of polyunsaturated fatty acids in plasma.

Conclusion These systemic approaches have highlighted networks of entities driving key biological processes involved in the phenotypic difference in feed efficiency between animals. Connecting transcriptomics and metabolic levels together had some additional benefits.

Keyword Feed efficiency, Fatty acids, Metabolomic, Molecular modules

5.2 Introduction

In the context of various geopolitical tensions and societal questions on the agri-agro-food systems, feed efficiency (FE) is a research priority to support food security and a sustainable meat production. Indeed, better FE is associated with a reduced amount of feeds needed for production and lower environmental wastes and emissions. Feed efficiency is measured on farms by the feed conversion ratio (`FCR`), an index calculated as the ratio of feed intake to body weight (`BW`) gain during a test period. Residual feed intake (`RFI`) has been also proposed as a refined measure of FE for genetic selection; it is calculated as the difference between observed feed intake and predicted feed intake from production and maintenance needs, which allows `RFI` to reflect digestive and metabolic variabilities [91, 92]. A number of studies have been performed in the past years to depict and understand the biological bases of FE. Based on the comparison of animals with low or high `FCR` or `RFI`, they all concluded to the complex nature of FE since

this trait integrates multiple biological pathways orchestrated in and by various tissues [92, 112, 113]. Among tissues, peripheral blood is a convenient and relatively easy sampling source of biological information that can highlight the variations in tissues metabolism and physiology to understand complex phenotypes [114] and with potential outcomes for applications in diagnostics and selection [115]. In pigs, about one thousand genes were found differentially expressed in the whole blood between two lines of pigs divergently selected for RFI [97, 116]. Gene set enrichment analysis on the whole blood transcriptome in beef cattle has also allowed identifying biological pathways associated with a divergent selection for low or high RFI [114]. Moreover, we recently showed that gene expression profiling in the whole blood is suitable to identify a few number of molecular candidate biomarkers for FCR in growing pigs, when gene expression levels were analyzed by machine learning algorithms based on classification or regression trees [93]. Likewise, metabolomic studies have shown that circulating concentrations of metabolites in the blood can be related to economically-important traits including FE [94]. However, all these approaches did not address the inter-individual variation in FE traits and did not attend to depict the interactions among the biological entities at a given level or different levels of omics organization. Therefore, a systemic approach considering the multiple relationships between molecules can add new insights in the architecture of complex traits such as FE.

Among various systems biology approaches, the weighted gene co-expression network analysis (WGCNA) has been proposed as a suitable method for defining interactions between transcripts of genes, by grouping them in modules of pairwise correlations to reveal the higher-order organization of the transcriptome [56]. Based on RNA-sequencing data in the liver, two co-expression networks were identified as associated with high or low FE in dairy cows [117]. Another approach based on linear models allowed to combine gene expression data and high throughput metabolomics data in skeletal muscle [118] ; in this study, pairs of metabolite-transcript associated with sphingolipid catabolism, multicellular organismal process, and purine metabolic processes were associated with differences in FE between two pig breeds and between two groups of pigs of low or high FE values.

The aim of the present study was to depict the biological bases underlying inter-individual variability in FE and other related traits in growing pigs by identifying small networks of interconnected gene transcripts in the whole blood and their relationships with global profiles of circulating metabolites. Eight molecular modules mainly composed of interconnected genes involved in immune and defense-related processes, were related to variability between pigs in FCR. Other important biological processes represented in these gene networks were the response to organic substance, ribosome biogenesis and translation, and cell development and

learning, respectively. One module of inter-connected expressed genes related to immune process was associated with circulating concentrations of omega3 fatty acids in the plasma, thus connecting transcripts to metabolites in the determinism of variability in FE.

5.3 Material and methods

5.3.1 Ethics

We reused transcriptomics and metabolomics datasets acquired in the whole blood from purebred French Large White pigs produced in a divergent selection experiment for RFI [97, 119], and that have been previously analyzed separately and for the line-associated effects only. The animal phenotypes have been published by [98]. These data were complemented by data on circulating concentrations of fatty acids (FA) in the plasma of the same pigs to specify lipid-related processes, that have not been previously published. In the original publications, the care and use of pigs were performed in compliance with the European Union legislation (directive 2010/63/EU). The protocol was approved by an Ethics Committee in Animal Experiment (Comité Rennais d'Ethique en matière d'Expérimentation Animale, CREEA N°007, agreement N°07–2012). All animals were reared and killed in compliance with the national regulations and according to procedures approved by the French veterinary Services. All methods were reported in accordance with ARRIVE guidelines. In the present study, reusing published data to perform new analyses for different animal traits perfectly fits with the 3R (Replacement, Reduction and Refinement) principles.

5.3.2 Origin of phenotypic data

Full description of the experimental design that provided the original datasets is referenced by [98]. Briefly, data were obtained from a total of 48 growing pigs (barrows) of two lines in the 7th generation of divergent selection for RFI, and fed diets formulated at isocaloric and isoproteic bases but differing in energy source and nutrients (lipids and fibers vs. starch), were tested from $74d \pm 0.3d$ of age to $132.5d \pm 0.5d$ (SEM) of age. All pigs were reared in isolated pens during the test period to allow the control of spontaneous feed intake, thus minimizing also the usual pen effect when pigs are reared in groups. From this publication, we considered body weight at slaughter (BW in kg), age at slaughter (in days), average daily feed intake (ADFI, in g/day), average daily gain (ADG, in g/day), FCR (calculated as the ratio between ADFI and

ADG during the feeding trial), and the percentage (relative to carcass weight) of the dorsal subcutaneous adipose tissue (%backfat) and of the loin muscle cut (%loin) as surrogates of body composition.

5.3.3 Transcriptomic dataset

The transcriptomic data were retrieved from NCBI's Gene Expression Omnibus (GEO) Subserie accession number GSE70838 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE70838). These data have been obtained from the whole blood collected in pigs at $132.5d \pm 0.5d$ (SEM) of age, by using a custom porcine microarray ($8 \times 60K$, GPL16524, Agilent Technologies France, Massy, France) containing $60,306$ porcine probes. Full description of the methods to produce the raw microarray data can be found in [97]. After quality filtration based on four criteria (background intensity value, diameter, saturation and uniformity of the spot), the original dataset contained $26,322$ annotated probes expressed in the whole blood. There were approximately 2.2 replicates per unique gene in the transcriptomic dataset (min: 1; max: 33).

5.3.4 Metabolomic dataset

The metabolomic dataset consisted in high resolution 1H-NMR spectra generated in plasma of the 48 pigs, and was retrieved from Jegou et al. [119]. The generated spectra were processed with one level of zero-filling and Fourier transformation after multiplying free induction decays by an exponential line broadening function of 0.3 Hz. The 1D NMR spectra were manually phase- and baseline-corrected, and referenced to the chemical shift of the alpha-glucose at delta 5.235. The bin area method was used to segment the spectra between 0.6 and 8.5 ppm using the intelligent variable size bucketing tool. Bin areas generated a matrix, which was normalized by dividing each integrated segment by the total area of the spectrum. This resulted in a new matrix that is used to perform statistical analyses. A total of 94 buckets, and consequently of 94 variables in this matrix, was considered for dedicated analyses in the present study. Buckets are individual metabolites or groups of two metabolites. They were notably assigned to different amino acids, creatine, lactate and glucose.

When using the 1H-NMR approach, lipids were all grouped as a single spectrum. In the present study, to investigate more deeply the role of the lipidic family in FE variability, the circulating concentrations of fatty acids (FA) in plasma were newly determined in the 48 pigs. For that, lipids were extracted from plasma as previously described for tissues [98], and methylated.

Dedicated analyses were then performed on a gas chromatograph (Nelson Analytical, Manchester, NH) equipped with a fused-silica capillary column ($30m \times 0.25mm$ internal diameter), with a base-deactivated silica stationary phase (a $0.25 - \mu m$ film thickness) filled with a stationary phase (80% biscyanopropyl and 20% cyanopropylphenyl) and using margaric acid (C17:0) as the internal standard. The furnace temperature was $180°C$, and injector and detector temperatures were $240°C$. Retention times and peak areas were determined. Peaks were identified by comparison with the retention times of standard FA methyl esters. Individual FAs were then quantified as percentages of the sum of FA identified in each sample. To facilitate the biological interpretation, FA have been then grouped into families of saturated FA with 14C or less (sC14:0), polyunsaturated FA of the omega-6 family (ssn-6c), and polyunsaturated FA of the omega-3 family (ssn-3). All other identified FA were kept as these. This led to a total of 14 variables representing circulating FA concentrations in plasma (Supplementary Table 1).

5.3.5 Construction of the weighted gene co-expression networks

Starting from a matrix whose individuals are pigs and features are probes expression levels (48 pigs $\times$ $26,322$ probes), we performed a hierarchical clustering to identify outlier individuals as recommended [120]. One pig was detected as an outlier and further removed from the dataset due to aberrant values. Then, we quantified the number of probes that were significantly linked to the animal phenotypic traits of interest (linear models with a $P\text{-}value < 0.05$ as cut-off). The results of the linear regressions showed that age at slaughter and line were the factors affecting the most the expression levels of molecular probes (19% of the probes significantly affected by age at slaughter and 14% by the line effect) (Supplementary Table 2). For next steps of analysis, we used the residuals of these linear regressions for the age effect, but preserved the intra- and inter-line variability of FE.

Corrected probe expression levels were then filtered so that the dataset can be analyzed by WGCNA as a single block, because "by block" comparison leads to ignore a large number of weak correlations. In accordance to WGCNA good practices, probes were filtered to remove non-varying ones, and only probes i whose log fold-change FC_i was greater than 1 were selected, with

$$log_2(FC_i) \; = \; max(log_2(expression_i)) - min(log_2(expression_i)) \tag{5.1}$$

The reduced dataset included $16,190$ probes for the 47 pigs.

To calculate the co-expression network, we used the Weighted Gene Co-expression Network

Analysis (WGCNA) step-by-step method [56], performed with R 4.2.2. The first step consisted in calculating a measure of co-expression similarity $s_{i,j}$ between each pair of probes to highlight the pairs of probes whose expression varies in a similar way. The adjacency matrix $A = [a_{i,j}]$ was then constructed by raising the co-expression similarity measure $s_{i,j}$ to the power β, using the `signed hybrid` method (only positive correlations were kept) [equation 5.2].

$$signed\ hybrid\ a_{ij} = \begin{cases} cor(x_i, x_j)^{\beta} & cor(x_i, x_j) > 0 \\ 0 & cor(x_i, x_j) < 0 \end{cases} \tag{5.2}$$

where a_{ij} is the element (i, j) of the adjacency matrix A, β is the soft-thresholding, x_i is the level expression of the ith probe, and $cor(x_i, x_j)$ is the Pearson correlation between expression profiles of the i and j probes.

β is a non-dichotomic soft thresholding that allows to evaluate connection between probes without losing the continuous character of the co-expression. Low correlations are better masked with high β values. We set β to 6 according to the criterion of the "approximate scale free topology" [121].

Considering probes as nodes, a weighted co-expression network can be then deduced from the adjacency matrix, by adding edges of weight a_{ij} between pairs of probes whose weight is strictly positive.

5.3.6 Detection of modules of co-expressed probes and their relationships with animal phenotypic traits

To detect modules in this network of co-expressed probes, a proximity measure was calculated using the Topological Overlap Measure [122], which is a valuable similarity measure set as default approach in WGCNA framework. The topological overlap of two nodes reflects their similarity in terms of the commonality of the nodes they connect to. Specifically, it calculates a correlation matrix from the expression data, calculates a soft threshold and assigns two genes a high topological overlap if they share common neighborhoods.

Then, an agglomerative hierarchical clustering was performed using the standard R function "hclust" and the unweighted average linkage criterion method (method = "average", UPGMA algorithm). With this linkage criterion, proximity between two clusters is the arithmetic mean of all the proximities between the probes of one, on one side, and the probes of the other, on the other side. This provides a hierarchical clustering leading to a dendrogram where each leaf corresponds to a probe and branches group together densely interconnected highly co-expressed

probes. Branch cutting was performed with the standard method for dynamic tree cut from the package dynamicTreeCut [56]. The sensitivity threshold was set to 2. To keep modules of highly co-expressed genes, we set the minimum module size to 25 probes. Each module was then summarized by the module eigengene ME, that is defined as the first principal component of a given module, i.e., a mathematical solution to condense the expression profile of the probes in the module. In the present study, we also checked that this first component always contributed to the majority - more than 40% - of the variance, and noticed that all other components explained each very small parts - less than 10% for the second component, etc.

A heatmap of correlations between ME and phenotypic traits was then produced. The heatmap can be examined to find the most significant associations. In this study, we considered $P \leq 0.05$ as significant association and $0.05 < P \leq 0.10$ as a tendency.

We also calculated the gene significance (`GS`) as (the absolute value of) the correlation between the probe and the phenotypic trait. For each module, we defined the quantitative measure of module membership (`MM`) as the correlation of ME and the gene expression profile. Using both the GS and MM measures, we can identify genes that have a high significance for FE traits (central players) and high module memberships in the modules. For a subset of modules, we provide a graphical representation of the sub-networks considering the annotated probes only. For that, pairs of probes with correlation coefficients greater than the 95th percentile in a given module were selected from the adjacency matrix, and represented using Cytoscape with nodes corresponding to probes and edges corresponding to the adjacency matrix values. For probes with low correlation with other probes, edges were represented in a different color.

5.3.7 Biological functional enrichment in modules of co-expressed probes

For modules that were significantly related to `FCR` based on the heatmap examination between module eigengenes (ME) and animal traits, we matched their constitutive molecular probes to the corresponding unique genes (official gene symbol), using the annotation provided by the manufacturer of the expression microarray. When applicable, the gene ontology (GO) terms for biological processes were then automatically searched in each module, using the Database for Annotation, Visualization and Integrated Discovery (DAVID) bioinformatics knowledgebase v2022q4 released (http://david.abcc.ncifcrf.gov/). The GOBP terms_FAT were selected to filter the broadest terms. The results were downloaded using the "Functional annotation clustering" option of the DAVID tool. Only clusters of terms with an enrichment score (measured by the geometric mean of the EASE score of all enriched annotations terms) greater

than 1.2 were considered. Within each cluster, the top GO term was listed together with its own enrichment score and the associated modified Fisher exact $P\text{-}value$.

5.3.8 Establishing profiles of circulating metabolites and evaluating connections between metabolic and transcriptomic levels

The second and third datasets considered in this study encompassed the circulating concentrations of metabolites and of FA, respectively. To reduce the dimension and facilitate correlation analyses with gene expression networks, these datasets were each summarized by Principal Component Analysis (PCA) using the R packages FactoMineR and factoextra [123, 124]. This would avoid bias when considering hundreds metabolic variables with only few modules of highly correlated genes. A PCA was used to summarize the profile in metabolites identified after 1H-NMR analysis (94 variables) and another PCA was used for circulating concentrations of FA (14 variables). From each table kept separately (one for 1H-NMR analysis and one for FA), PCA transforms the original (mean-centered) observations to a new set of variables (dimensions) using the eigenvectors and eigenvalues calculated from a covariance matrix of the original variables. The first components of the PCA were called Dim$i$_Metab for the metabolomic 1H-NMR table and Dim$i$_FA for the FA table, respectively, with i = 1 to 5. These dimensions were linear combinations of the original variables. There is no consensus on the best number of dimensions to be considered in PCA analysis, and this depends on the objectives. It is generally assumed keeping dimensions to reach 80% or more of the variance, but the inflation of dimensions is also not recommended. In this study, we first examined the distribution of the variance explained by each of the tenth first dimensions in each PCA, and based on this, we kept the first five dimensions that explained around 80% of the total variance.

To connect information at the two omics levels, the dimensions of each PCA were then correlated to the eigengenes of the WGCNA modules (ME) by using Pearson correlations. The correlations were represented by heatmaps to facilitate the description.

5.4 Results

Data obtained in a total of 47 growing pigs with inter-individual differences in `FCR` (i.e., the measure of FE on farms) due to genetic selection for `RFI` and to the diet received during a test period of 58 days, were considered in the present study. In addition to `FCR`, the average daily gain (`ADG`) and average daily feed intake (`ADFI`) (i.e., the two components of `FCR`), and body

composition estimated by percentage (relative to carcass weight) of backfat (`%backfat`) and of the loin cut (`%loin`) were also obtained (Supplementary Table 1).

5.4.1 Definition of gene co-expression network in the whole blood of pigs

A network was built with the WGCNA package from the expression levels of $16,190$ molecular probes expressed in the whole blood, where nodes correspond to the expression profile of the molecular probes, and edges are determined by the pairwise correlations between probes expression (the adjacency matrix is available at https://data-access.cesgo.org/index.php/s/YPz0J2ItxIEuN5M). This network was then analysed to find modules defined as groups of co-expressed probes that may represent the molecular architecture behind the animal phenotypic traits.

After excluding the grey module which is used to hold all the probes that do not clearly belong to any other modules i.e., probes that are not co-expressed, we show that the co-expression network was composed of 33 modules composed of 27 to $3,829$ molecular probes (Supplementary Table 3). Annotations were used when applicable to identify the corresponding genes. The distribution in the number of probes and their corresponding unique genes per module is shown in Figure 5.1.

The module eigengene (ME) was then calculated as the representative of expression profiles of the genes in the module. The module membership (MM) was calculated by correlating ME and the expression profile of the probes within each module. By definition, the closer to 1 or -1 is MM, the higher is the gene connected to ME. The medians of the MM values indicated a satisfactory clustering from the whole network of the molecular probes expressed in the whole blood of the 47 pigs. The values are available in Supplementary Table 3.

5.4.2 Relationships between modules of co-expressed genes and animal phenotypic traits

The heatmap of the correlation coefficients between the module eigengene (ME) of each WGCNA module and the animal phenotypic traits is presented in Figure 5.1. For six modules, the ME was significantly correlated ($P\text{-}value \leq 0.05$) with `FCR`: they were the modules `violet` (probes: 32, unique genes: 16, $P\text{-}value = 8e-04$), `darkred` (probes: 82, unique genes: 53, $P\text{-}value = 2e-04$), `royalblue` (probes: 111, unique genes: 81, P-

$value = 0.04$), `lightcyan` (probes: 180, unique genes: 114, $P\text{-}value = 0.008$), `white` (probes: 59, unique genes: 28, $P\text{-}value = 0.04$), and `darkorange` (probes: 60, unique genes: 21, $P\text{-}value = 0.05$). For two other modules, ME tended ($P\text{-}value < 0.1$) to be correlated with `FCR`: they were the modules `darkolivegreen` (probes: 27, unique genes: 13, $P\text{-}value = 0.08$) and `steelblue` (probes: 45, unique genes: 23, $P\text{-}value = 0.09$).

For the modules `violet`, `darkred`, `royalblue`, `lightcyan`, and `white`, the correlations between ME and `ADG` were also significant, and for the module `darkorange`, there was a trend for correlation between ME and `ADG`. As expected, the signs of correlation between ME and `ADG` and between ME and `FCR` (which is the ratio between `ADFI` and `ADG` during the test period), were opposite. For the modules `darkolivegreen` and `steelblue`, the correlation coefficients between ME and `ADG` did not reach statistical significance.

Four of the eight modules associated with `FCR` also displayed significant correlations with `%loin` (with opposite signs of correlation): the modules `violet`, `darkred`, `lightcyan` and `white`. For these modules, there was no significant correlation between ME and `%backfat`.

Finally, none of the eight modules associated with `FCR` were significantly related to `ADFI`. In the whole network, only the `darkgreen` module (probes: 82, unique genes: 43, $P\text{-}value = 8e-04$) was highly correlated with `ADFI`; it was also significantly related with `%loin` ($P\text{-}value = 0.004$) and `%backfat` ($P\text{-}value = 2e-04$). The `saddlebrown` module (probes: 46, unique genes: 12) tended to be correlated with `ADFI` ($P\text{-}value = 0.1$) and with `%backfat` ($P\text{-}value = 0.1$). Finally, the eigengenes of the modules `green` (probes: $1,113$, unique genes: 412, $P\text{-}value = 0.07$), `brown` (probes: $1,441$, unique genes: 773, $P\text{-}value = 0.02$) and `skyblue` (probes: 46, unique genes: 45, $P\text{-}value = 0.05$) displayed significant associations with `%loin`, but without significant correlations with any other animal phenotypic traits.

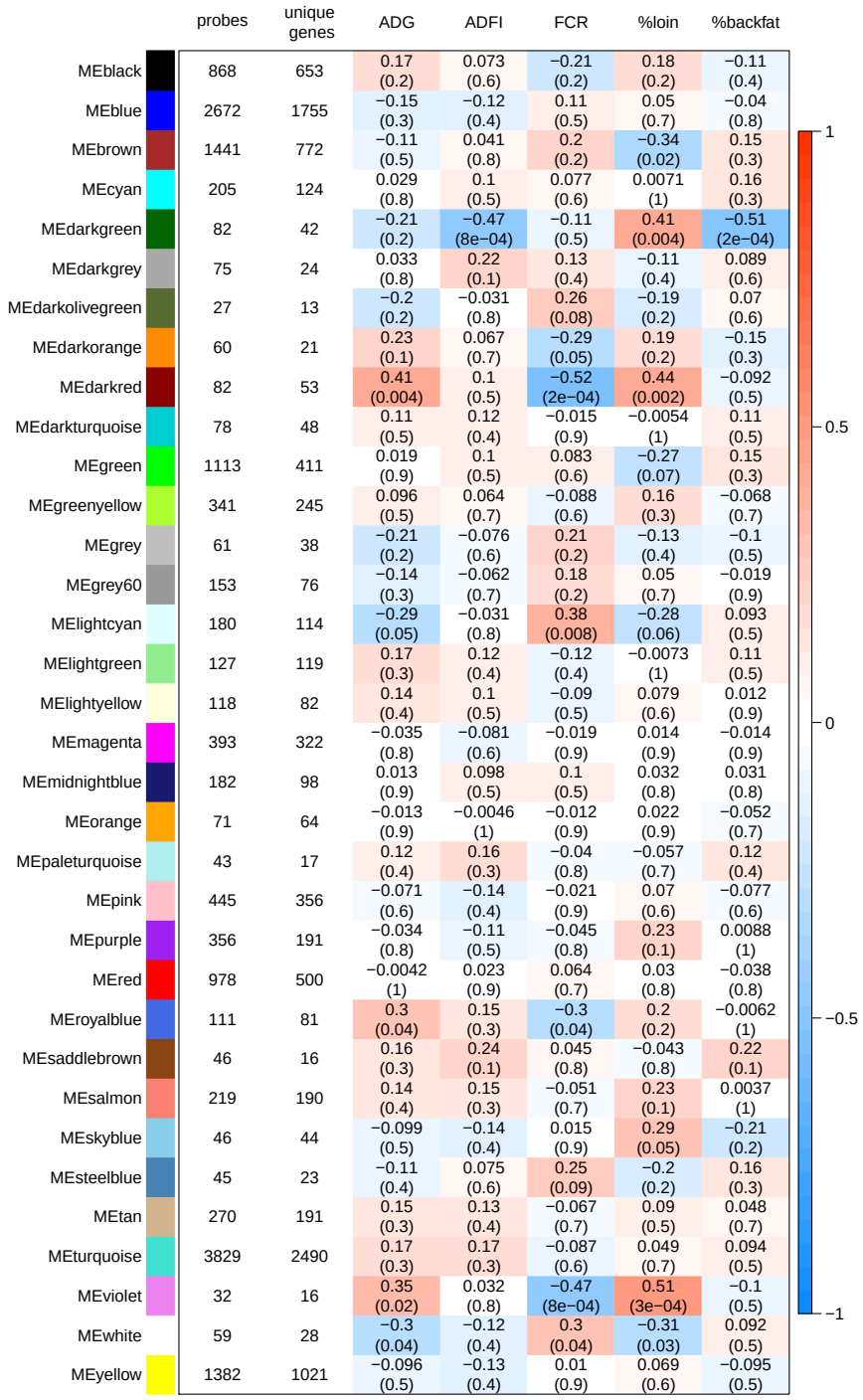
probes
unique genes
ADG
ADFI
FCR
%loin
%backfat
MEblack 868 653 0.17 (0.2) 0.073 (0.6) −0.21 (0.2) 0.18 (0.2) −0.11 (0.4)
MEblue 2672 1755 −0.15 (0.3) −0.12 (0.4) 0.11 (0.5) 0.05 (0.7) −0.04 (0.8)
MEbrown 1441 772 −0.11 (0.5) 0.041 (0.8) 0.2 (0.2) −0.34 (0.02) 0.15 (0.3)
MEcyan 205 124 0.029 (0.8) 0.1 (0.5) 0.077 (0.6) 0.0071 (1) 0.16 (0.3)
MEdarkgreen 82 42 −0.21 (0.2) −0.47 (8e−04) −0.11 (0.5) 0.41 (0.004) −0.51 (2e−04)
MEdarkgrey 75 24 0.033 (0.8) 0.22 (0.1) 0.13 (0.4) −0.11 (0.4) 0.089 (0.6)
MEdarkolivegreen 27 13 −0.2 (0.2) −0.031 (0.8) 0.26 (0.08) −0.19 (0.2) 0.07 (0.6)
MEdarkorange 60 21 0.23 (0.1) 0.067 (0.7) −0.29 (0.05) 0.19 (0.2) −0.15 (0.3)
MEdarkred 82 53 0.41 (0.004) 0.1 (0.5) −0.52 (2e−04) 0.44 (0.002) −0.092 (0.5)
MEdarkturquoise 78 48 0.11 (0.5) 0.12 (0.4) −0.015 (0.9) −0.0054 (1) 0.11 (0.5)
MEgreen 1113 411 0.019 (0.9) 0.1 (0.5) 0.083 (0.6) −0.27 (0.07) 0.15 (0.3)
MEgreenyellow 341 245 0.096 (0.5) 0.064 (0.7) −0.088 (0.6) 0.16 (0.3) −0.068 (0.7)
MEgrey 61 38 −0.21 (0.2) −0.076 (0.6) 0.21 (0.2) −0.13 (0.4) −0.1 (0.5)
MEgrey60 153 76 −0.14 (0.3) −0.062 (0.7) 0.18 (0.2) 0.05 (0.7) −0.019 (0.9)
MElightcyan 180 114 −0.29 (0.05) −0.031 (0.8) 0.38 (0.008) −0.28 (0.06) 0.093 (0.5)
MElightgreen 127 119 0.17 (0.3) 0.12 (0.4) −0.12 (0.4) −0.0073 (1) 0.11 (0.5)
MElightyellow 118 82 0.14 (0.4) 0.1 (0.5) −0.09 (0.5) 0.079 (0.6) 0.012 (0.9)
MEmagenta 393 322 −0.035 (0.8) −0.081 (0.6) −0.019 (0.9) 0.014 (0.9) −0.014 (0.9)
MEmidnightblue 182 98 0.013 (0.9) 0.098 (0.5) 0.1 (0.5) 0.032 (0.8) 0.031 (0.8)
MEorange 71 64 −0.013 (0.9) −0.0046 (1) −0.012 (0.9) 0.022 (0.9) −0.052 (0.7)
MEpaleturquoise 43 17 0.12 (0.4) 0.16 (0.3) −0.04 (0.8) −0.057 (0.7) 0.12 (0.4)
MEpink 445 356 −0.071 (0.6) −0.14 (0.4) −0.021 (0.9) 0.07 (0.6) −0.077 (0.6)
MEpurple 356 191 −0.034 (0.8) −0.11 (0.5) −0.045 (0.8) 0.23 (0.1) 0.0088 (1)
MEred 978 500 −0.0042 (1) 0.023 (0.9) 0.064 (0.7) 0.03 (0.8) −0.038 (0.8)
MEroyalblue 111 81 0.3 (0.04) 0.15 (0.3) −0.3 (0.04) 0.2 (0.2) −0.0062 (1)
MEsaddlebrown 46 16 0.16 (0.3) 0.24 (0.1) 0.045 (0.8) −0.043 (0.8) 0.22 (0.1)
MEsalmon 219 190 0.14 (0.4) 0.15 (0.3) −0.051 (0.7) 0.23 (0.1) 0.0037 (1)
MEskyblue 46 44 −0.099 (0.5) −0.14 (0.4) 0.015 (0.9) 0.29 (0.05) −0.21 (0.2)
MEsteelblue 45 23 −0.11 (0.4) 0.075 (0.6) 0.25 (0.09) −0.2 (0.2) 0.16 (0.3)
MEtan 270 191 0.15 (0.3) 0.13 (0.4) −0.067 (0.7) 0.09 (0.5) 0.048 (0.7)
MEturquoise 3829 2490 0.17 (0.3) 0.17 (0.3) −0.087 (0.6) 0.049 (0.7) 0.094 (0.5)
MEviolet 32 16 0.35 (0.02) 0.032 (0.8) −0.47 (8e−04) 0.51 (3e−04) −0.1 (0.5)
MEwhite 59 28 −0.3 (0.04) −0.12 (0.4) 0.3 (0.04) −0.31 (0.03) 0.092 (0.5)
MEyellow 1382 1021 −0.096 (0.5) −0.13 (0.4) 0.01 (0.9) 0.069 (0.6) −0.095 (0.5)
1
0.5
0
−0.5
−1

Figure 5.1 – **Heatmap of correlations between module eigengenes and animal phenotypic traits.** Module eigengene (ME) was the representative of gene expression profile in the module of co-expressed molecular probes elicited from the weighted gene correlated network analysis (WGCNA) from microarray data in the whole blood of 47 growing pigs. Animal phenotypic traits were recorded during a test period of 58 days. The heatmap indicates the Pearson correlation coefficient between ME and the phenotypic trait together with the statistical significance ($Pvalue$).
Abbreviations: `ADG` = average daily gain; `ADFI` = average daily feed intake; `FCR` = feed conversion ratio; `%loin` = percentage of loin weight relative to carcass weight; `%backfat` = percentage of dorsal subcutaneous fat tissue weight relative to carcass weight; ME = module eigengene

5.4.3 Close-vicinity of the different modules of co-expressed genes

To evaluate the connectivity between modules in the network, a hierarchical clustering was performed between the eigengenes (ME) of the modules. The resulting dendrogram is shown in Figure 5.2, in which the eight modules that were significantly associated or tending to be associated with `FCR` are enlightened. Among these eight modules, two clusters were identified. The first cluster associated the `lightcyan` (114 genes), `steelblue` (23 genes) and `darkolivegreen` (13 genes) modules. The second cluster associated the `darkred` (53 genes), `violet` (16 genes) and `royalblue` (81 genes) modules. The `white` (28 genes) and the `darkorange` (21 genes) modules were isolated in the dendrogram. In addition, the `green` (411 genes) and `brown` (772 genes) modules, that were significantly associated with `%loin`, were clustered together (Figure 5.2).

5.4.4 Functional enrichment of the modules in biological processes

For the eight modules identified as significantly correlated or tending to be correlated with `FCR`, an enrichment analysis was performed to find the main biological processes shared by the co-expressed gene transcripts within each module (Table 5.1). Annotations of the probes were first retrieved, and the corresponding gene name was associated to each probe when applicable. The DAVID tool was used on the gene list uploaded for each module.

The `lightcyan` which was the biggest module correlated to `FCR` (180 probes corresponding to 114 unique genes), displayed a large number of different clusters of biological processes. On the opposite, for the other seven modules considered, there was/were only one to three clusters ($E > 1$) identified among the GO terms in each module (Table 5.1). This indicates a good consistency of the biological processes shared by the intra-connected genes within each module. The modules `violet`, `royalblue`, `darkorange`, `lightcyan` and `darkolivegreen`

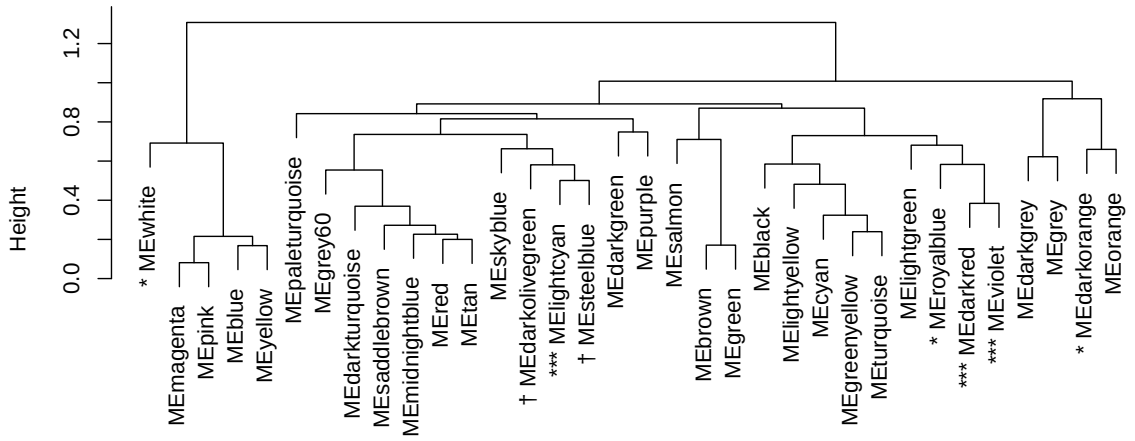

Figure 5.2 – **Hierarchical clustering of module eigengenes.** The modules that were found highly correlated with feed conversion ratio (`FCR`) are enlightened (*** $P \leq 0.001$, * $P \leq 0.05$, and † $0.05 < P \leq 0.10$).

showed a predominance of immune, inflammatory and defense-related pathways across their constitutive genes. The `darkorange` module also included genes involved in the response to organic substance. The `darkred` module was rather oriented towards ribosome biogenesis and the process regulating translation. The `white` module was related to circulatory cell development and to learning.

The biological processes identified in the other modules related to feed intake (`ADFI`) or body composition (`%loin` and `%backfat`), but not to `FCR`, were described in Table 5.2.

The module `saddlebrown` that tended to be correlated to `ADFI` and `%backfat` was composed of 16 unique genes participating to the response to stimulus. The `skyblue` module that was significantly correlated to muscle mass (`%loin`), showed a predominance of genes related to protein metabolism among its 44 unique constitutive genes (Table 5.2). The `green` module was a large module of 411 unique genes related to various processes such as epigenetics processes (chromatin organization, histone methylation), to cellular responses to compounds (nitrogen, organic cyclic) for stress and defense, and to circadian regulation. The `brown` was also a big module of 772 unique genes that regulated cell differentiation, protein and lipid processes, and signaling pathways.

Table 5.1 – **Functional enrichment in biological pathways for molecular modules related to FCR.**
The DAVID tool was used to identify the top enriched pathways across the list of unique annotated genes within each module. The gene ontology (GO) terms for biological processes are indicated together with enrichment score (E) of the process and Fisher P value.

Module	**probes**	**unique genes**	**match ID DAVID**	**GO_terms**
darkolivegreen	27	13	9	GO:0006954 inflammatory response $E = 15.6$ $P < 0.05$
darkorange	60	21	15	GO:0010033 response to organic substance $E = 4.4$ $P < 0.001$, GO:0002376 immune system process $E = 2.9$ $P < 0.05$
darkred	82	53	44	GO:0042273 ribosomal large subunit biogenesis $E = 18.8$ $P < 0.01$, GO:0002181 cytoplasmic translation $E = 16.5$ $P < 0.01$
lightcyan	180	114	96	GO:0007166 cell surface receptor signaling pathway $E = 2.3$ $P < 0.001$, GO:0032502 developmental process $E = 1.5$ $P < 0.001$, GO:0006955 immune response $E = 2.74$ $P < 0.001$, GO:0006909 phagocytosis $E = 4.5$ $P < 0.01$, GO:0042113 B cell activation $E = 9.8$ $P < 0.001$, GO:0070887 cellular response to chemical stimulus $E = 1.7$ $P < 0.01$, GO:0045088 regulation of innate immune response $E = 3.4$ $P < 0.05$, GO:0050727 regulation of inflammatory response $E = 3.1$ $P < 0.05$
royalblue	111	81	71	GO:0050852 T cell receptor signaling pathway $E = 21.1$ $P < 0.001$, GO:0030155 regulation of cell adhesion $E = 5.6$ $P < 0.001$, GO:0048869 cellular developmental process $E = 2.0$ $P < 0.001$
steelblue	45	23	10	GO:0008104 protein localization $E = 3.9$ $P < 0.05$
violet	32	16	12	GO 0045087 innate immune response $E = 9.4$ $P < 0.001$, GO:0016567 protein ubiquitination $E = 7.5$ $P < 0.05$, GO:0006952 defense response $E = 6.0$ $P < 0.001$
white	59	28	23	GO:0072359 circulatory system development $E = 4.4$ $P < 0.01$, GO:0007612 learning $E = 16.9$ $P < 0.01$, GO:0030036 actin cytoskeleton organization $E = 5.0$ $P < 0.05$

Table 5.2 – **Functional enrichment of molecular modules correlated with feed intake or body composition, but not to `FCR`.**
The DAVID tool was used to identify the top enriched pathways across the list of unique annotated genes within each module. The gene ontology (GO) terms for biological processes are indicated together with enrichment score (E) of the process and Fisher P value.

Module	probes	unique genes	match ID DAVID	GO_terms
brown	1,441	772	623	GO:0009966 regulation of signal transduction $E = 1.4$ $P < 0.001$, GO:0050790 regulation of catalytic activity $E = 1.4$ $P < 0.001$, GO:0007169 transmembrane receptor protein tyrosine kinase signaling pathway $E = 1.9$ $P < 0.001$, GO:0036211 protein modification process $E = 1.3$ $P < 0.001$, GO:0042886 amide transport $E = 1.4$ $P < 0.001$, GO:0007167 enzyme linked receptor protein signaling pathway $E = 1.7$ $P < 0.001$, GO:0031400 negative regulation of protein modification process $E = 1.9$ $P < 0.001$, GO:0042692 muscle cell differentiation $E = 1.8$ $P < 0.01$, GO:0071363 cellular response to growth factor stimulus $E = 1.9$ $P < 0.001$, GO:0031331 positive regulation of cellular catabolic process $E = 1.7$ $P < 0.05$, GO:0008654 phospholipid biosynthetic process $E = 2.5$ $P < 0.01$, GO:1903320 regulation of protein modification by small protein conjugation or removal $E = 2.2$ $P < 0.01$, GO:0045595 regulation of cell differentiation $E = 1.4$ $P < 0.01$, GO:0034284 response to monosaccharide $E = 2.1$ $P < 0.01$
darkgreen	82	42	33	GO:0009636 response to toxic substance $E = 9.9$ $P < 0.01$, GO:0032496 response to lipopolysaccharide $E = 7.9$ $P < 0.01$, GO:0051128 regulation of cellular component organization $E = 2.2$ $P < 0.05$, GO:0042063 gliogenesis $E = 8.9$ $P < 0.01$
green	1,113	411	356	GO:0006325 chromatin organization $E = 2.0$ $P < 0.001$, GO:0009891 positive regulation of biosynthetic process $E = 1.7$ $P < 0.001$, GO:0034968 histone lysine methylation $E = 5.3$ $P < 0.001$, GO:0033555 multicellular organismal response to stress $E = 4.8$ $P < 0.01$, GO:1901699 cellular response to nitrogen compound $E = 2.1$ $P < 0.001$, GO:0032922 circadian regulation of gene expression $E = 5.7$ $P < 0.01$, GO:0006622 protein targeting to lysosome $E = 9.37$ $P < 0.001$, GO:0016050 vesicle organization $E = 2.4$ $P < 0.001$, GO:0071407 cellular response to organic cyclic compound $E = 2.0$ $P < 0.01$, GO:0016071 mRNA metabolic process $E = 2.0$ $P < 0.001$, GO:0009611 response to wounding $E = 2.1$ $P < 0.01$
saddlebrown	46	16	13	GO:0006357 regulation of transcription $E = 3.72$ $P < 0.01$, GO:0009628 response to abiotic stimulus $E = 9.6$ $P < 0.001$
skyblue	46	44	36	GO:0006807 nitrogen compound metabolic process $E = 1.80$ $P < 0.001$, GO:0045184 establishment of protein localization $E = 2.7$ $P < 0.01$, GO:0034660 ncRNA metabolic process $E = 4.8$ $P < 0.05$

5.4.5 Hierarchy of expressed genes in the modules related to feed efficiency traits

To determine which expressed genes accounted the most in the correlations between the module eigengene (ME) and FCR, we calculated the Gene Significance (GS) in each module, and expressed the GS as a function of FCR. The top genes are listed in Table 5.3 and the all data are provided in Supplementary Table 3.

For the white module, IGDCC3, TMEM14C, HRH1, HTR7 and AFF1 were notably found as top genes. For the violet module, SLPI, P2RY1, MUCA, MED8, HSPA1B and HSP70.2 were the most important genes triggering the correlation of the module with FCR. For the royalblue module, POFUT1, LPAR3, CCR7, PTTG1, STRN, NPY and PPAR26 were pointed as important in the correlation with FCR. For the darkred module, EIF1B, RPL14 and KRTCAP3 were among the top 15 genes. For the lightcyan module, ITGAD, DDG, HTRA1, EBF1 and CBR3 were notably listed. A graphical representation of the importance of the genes in the modules is also provided. For the royalblue (Figure 5.3), this shows that the probes that were highly correlated to other probes were listed in the top list of probes based on their module membership (MM).

Table 5.3 – **Top genes in the molecular modules related to FCR.**
The unique genes corresponding to the annotated probes were listed in each module according to their GS.FCR value. In the table, only genes with a value greater than 0.3 for GS.FCR are indicated. The GS.FCR is the correlation between Gene Significance (GS) of the module eigengene and FCR.

darkolivegreen	darkorange	darkred	lightcyan	royalblue	steelblue	violet	white
PTGER3	BMP6	SLCO2B1	ITGAD	POFUT1	UQCC2	SLPI	IGDCC3
TUBB6	PADI2	E4	DDC	LPAR3		P2RY1	TMEM14C
SLA-DRB1	DNAJB9	CCR3	HTRA1	CCR7		MUC4	KIAA0247
	WBSCR27	SLC46A2	EBF1	PTTG1		MED8	HRH1
	TAOK3	EIF1B	CBR3	NIPSNAP3B		HSPA1B	HTR7
	SERHL2	ZFAND6	MYBL1	EPHB6		TR10D	AFF1
		CD300C	C4BPB	SLC4A11		HSP70.2	ZUFSP
		FAM102A	KIAA0556	STRN			LDB3
		RPL14	MS4A1	ZCCHC10			PROX1
		WWP1	SLA-DOA	FMNL3			ACVR2B
		KRTCAP3	HMG20A	SKAP1			DGKA
		POLB	CYAC3	NPY			LUM
		PIGL	RALGPS2	PPP1R26			TBC1D19
		TMEM52B	ELL3	IZUMO4			
		PLA2G12A	PIKFYVE	NIPSNAP3A			

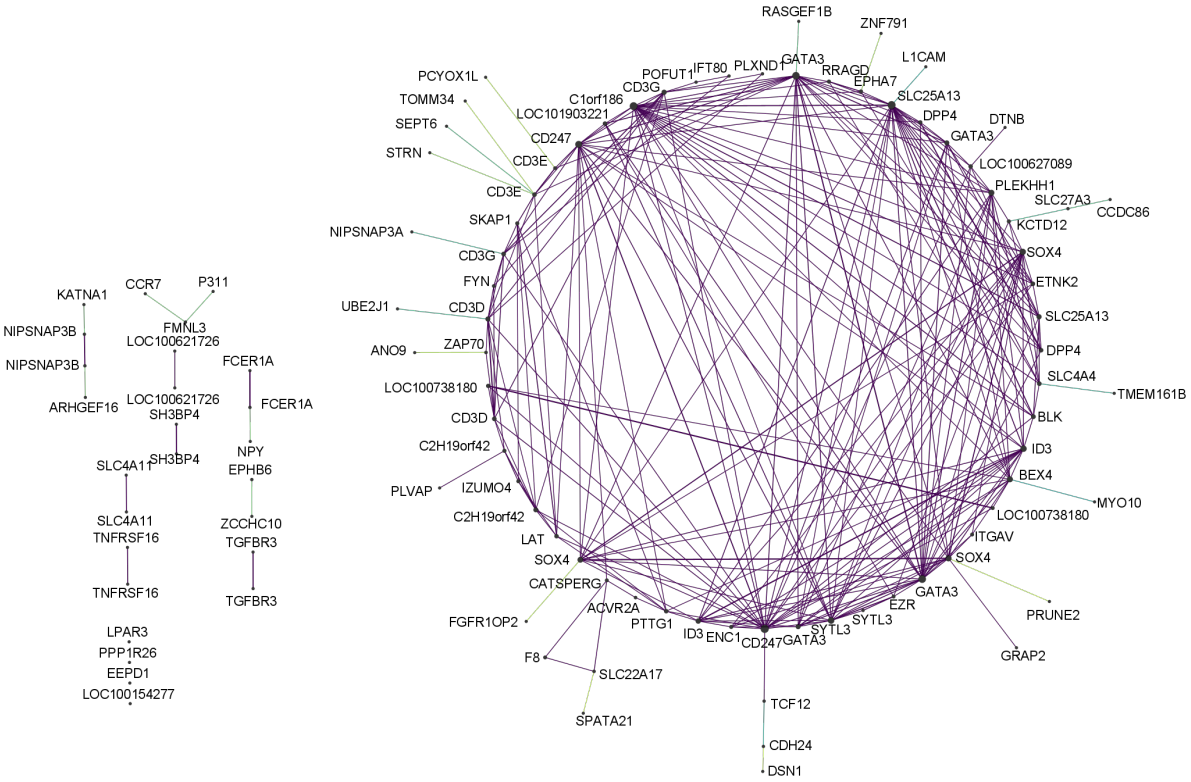

Figure 5.3 – **Graphical representation of the `royalbue` module.** The network was constructed from the adjacency matrix of the `royalblue` module using Cytoscape. The nodes are the molecular probes, labeled with their annotation and connected by purple edges that represent the correlation coefficients greater than the $95th$ percentile and by green edges, for probes that were not sufficiently correlated to other probes, to the annotated probe to which they were the most correlated. The size of the nodes is a function of their degree.

5.4.6 Metabolic profiles in the whole blood

The second part of the present study addressed the metabolic level by considering several variables obtained in the circulating blood of the 47 pigs, after non-target analysis (1H-NMR spectra) of plasma and specific analysis of the lipidic fraction (gas chromatography-derived information) of plasma. The relationships between the 94 variables obtained from the 1H-NMR spectra were summarized by the first five dimensions of a principal component analysis (PCA). We showed that the two first dimensions represented 41.4% and 26.8% of the total variability, respectively. Figure 5.4 shows the corresponding correlation circle. These two dimensions fitted with the objective of summarizing metabolic profiles across the pigs. The first dimension

of this PCA mainly opposed lactate to the majority of the identified amino acids (lysine, tyrosine, valine, phenylalanine, methionine, leucine, isoleucine, glutamine-glutamate) and to high density lipoproteins (HDL). The second dimension mainly opposed glucose, eventually combined with other molecules, to circulating lipoproteins (very low density lipoproteins VLDL, low density lipoprotein LDL, and lipids) and to threonine. In addition, because specific metabolites in plasma are considered as end-points in biochemical reactions, the third, fourth and fifth dimensions were also considered in an exploratory factor analysis of few groups of variables that may be helpful to generate biological theory. The third, fourth and fifth dimensions represented respectively 7.9%, 6.0% and 4.1% of the total variability. The third dimension mainly opposed circulating concentrations of glutamine (Gln), glutamate (Glu) and proline (Pro) on one hand, and beta-hydroxybutyrate on the other hand. The fifth dimension opposed circulating concentrations of betaine (bet) and trimethylamine N-oxide (TMAO; i.e., a metabolite produced by the liver and associated with microbiote metabolism), to VLDL and inositol concentrations. Correlation circles for the third, fourth and fifth are available in Supplementary Figures 1 and 2. Overall, the first five dimensions explained 86% of the variance.

Considering specifically the lipid fraction of the plasma, we analyzed the fatty acids (FA) composition by target methodology. From the 30 FA (10 to 22 carbon chains) that can be analyzed, some of them were present in negligible concentrations in the plasma or even cannot be detected from the background in some pigs (e.g.; C10:0, C12:0, C18:4 n-3). Therefore, parts of the FA were grouped in biologically relevant families (saturated FA with 14 carbons or less, n-6 FA family; n-3 FA family). This led to a total of 14 variables representing single FA or groups of FA. They were then represented by a second PCA to summarize the profiles in circulating FA among the 47 pigs. Figure 5.5 shows the corresponding correlation circle. The first dimension represented 42.5% of the total variability and opposed omega-6 (n-6) family of FA and to a lesser extent omega-3 (n-3) FA, to saturated family of FA. The second dimension represented 13.5% of the total variability and opposed C15:0 to C20:0 FA. Similarly to what was used for metabolomics, we also considered the third, fourth and fifth dimensions, that represented respectively 8.0%, 7.5% and 7.1% of the total variability, respectively. The third dimension opposed C20:1 to C20:2 FA on one hand and C20:0 FA on the other hand, whereas the fourth dimension opposed the sum of n-3 FA to C20:1 FA. Correlation circles for the third, fourth and fifth are available in Supplementary Figures 3 and 4. Overall, the first five dimensions explained 79% of the variance.

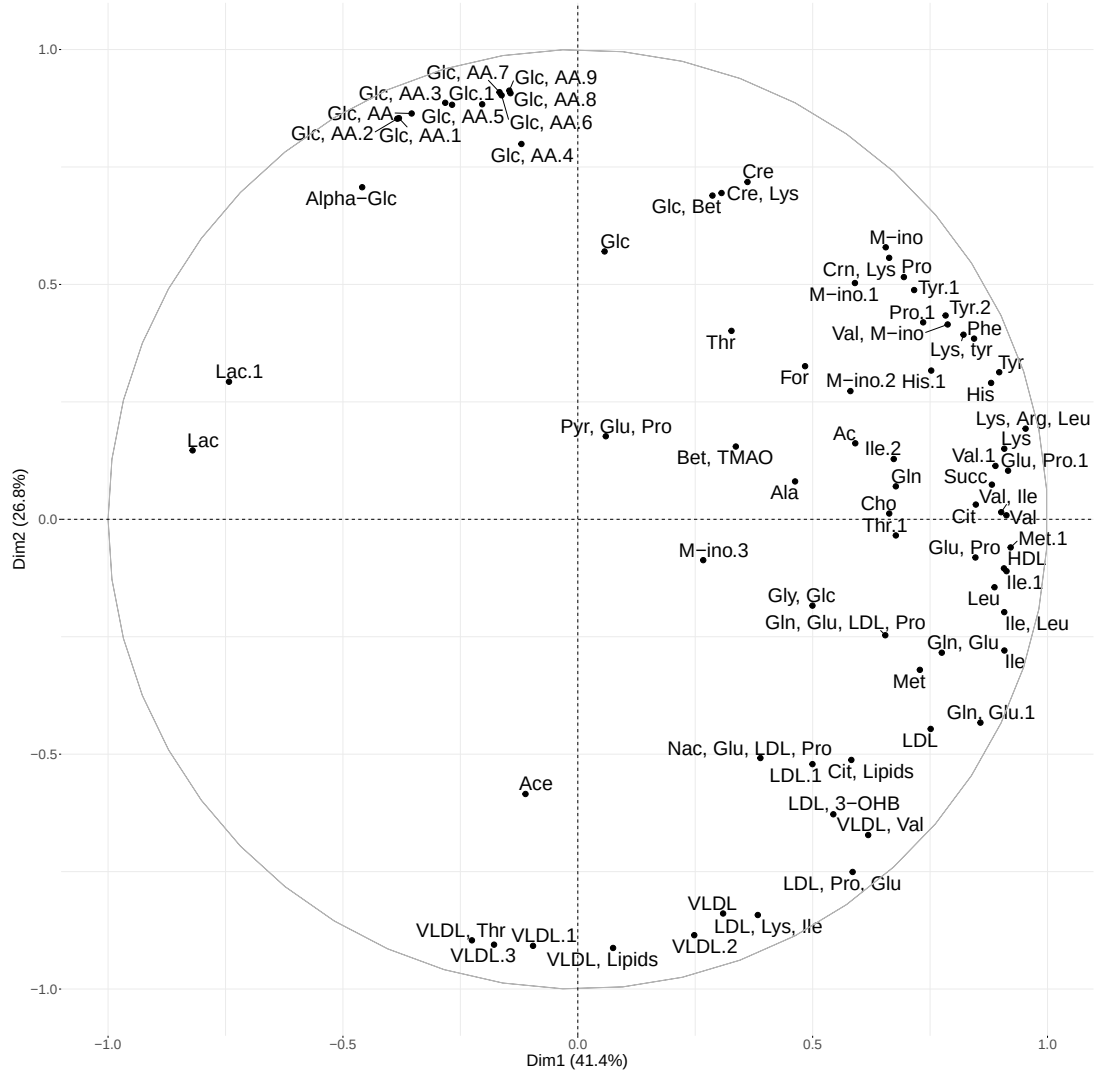

Figure 5.4 – **Correlation circle of the principal component analysis summarizing the profiles of circulating metabolites.** 1H-NMR spectra were obtained in the plasma prepared from the whole blood of 47 growing pigs. The matrix of correlations was calculated from 94 individual variables corresponding to the different annotated spectra.

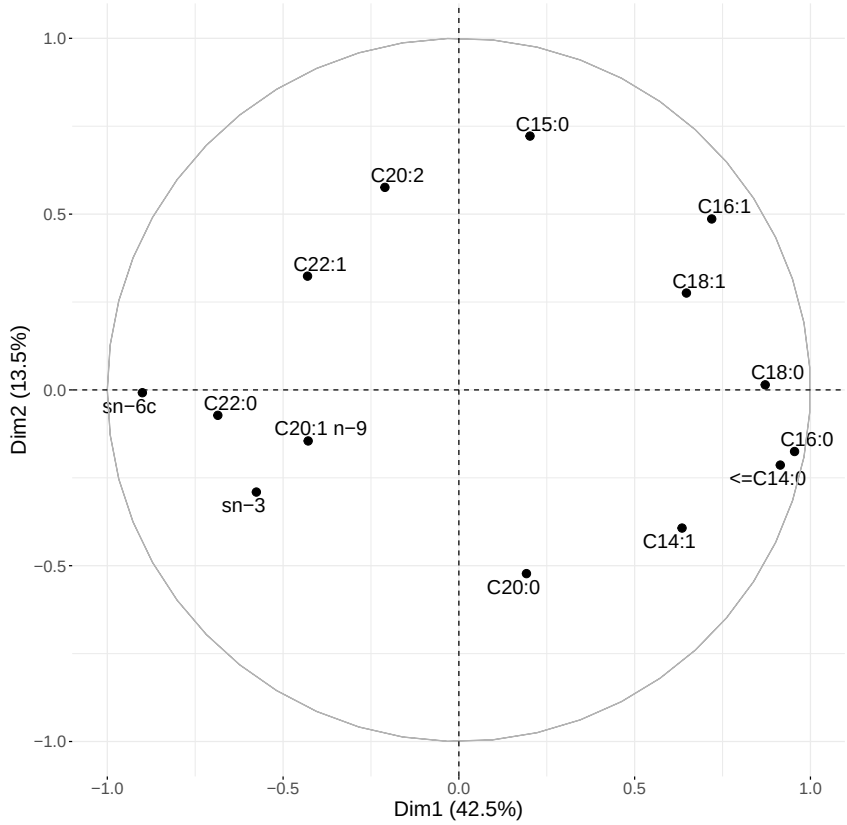

Figure 5.5 – **Correlation circle of the principal component analysis summarizing fatty acid composition in blood.** The fatty acid composition (in percentage) was obtained in the plasma prepared from the whole blood of 47 growing pigs by using gas chromatography. Some of the individual FA were grouped in biologically relevant families (saturated FA with 14 carbons or less, omega-6 sum of $n-6$ and omega-3 sum of $n-3$), whereas the other fatty acids were kept as these.

Table 5.4 – **Relationships between transcriptomic and metabolomic levels.**
Modules of co-expressed genes were identified from microarray data in the whole blood by using weighted gene correlation network analysis (WGCNA). Circulating biochemical molecules in the plasma were analysed by 1H-NMR (metabolites; met) or target gas chromatography for the lipid fraction subset (FA) and the data were summarized by weighted linear regressions (dim) using principal component analysis (PCA). Correlations were calculated between the module eigengenes (ME) and the first five dimensions of each PCA. The table indicates the number of significant correlations or trends between modules of coexpressed genes and PCA dimensions.

Dimensions of the principal component analysis of the metabolic level in blood					
Modules	Dim1_met	Dim2_met	Dim3_met	Dim4_met	Dim5_met
Significantly correlated ($Pvalue \leq 0.05$)	1	2	8	0	11
Trend ($Pvalue \leq 0.1$)	3	5	6	7	6
Dimensions of the principal component analysis of the fatty acids in blood					
Modules	Dim1_FA	Dim2_FA	Dim3_FA	Dim4_FA	Dim5_FA
Significantly correlated ($Pvalue \leq 0.05$)	2	1	1	1	1
Trend ($Pvalue \leq 0.1$)	4	0	3	3	0

5.4.7 Connecting the two omics levels

We calculated the correlations between the eigengenes of the molecular modules (ME) and the profiles in circulating metabolites or fatty acids represented by the different dimensions of each PCA. This allows connecting the transcriptome (*via* the WGCNA modules) and the metabolome (*via* the principal components of the PCA).

The numbers of modules for which ME was correlated with at least one of the five dimensions of each PCA are presented in Table 5.4. A heatmap representing the correlation coefficients calculated between ME of all modules identified from microarrray data and the first five dimensions of the PCA, is available in Supplementary Figures 5 and 6.

There were few associations between molecular modules and metabolic and lipid profiles. A summary is presented in Figure 5.6 considering only the list of modules of interconnected genes that have been found to be associated with the phenotypic traits of interest in the previous subsection Definition of gene co-expression network in the whole blood of pigs. There were no significant correlations between the eigengenes of the modules associated with `FCR` and the profiles in circulating metabolites. Only the `darkorange` module tended to be associated with the second dimension (dim2_met) which opposed circulating concentrations of glucose to circulating concentrations of LDL and VLDL lipoproteins. For the other animal traits, the `darkgreen` module which was significantly associated with `ADFI`, was highly correlated

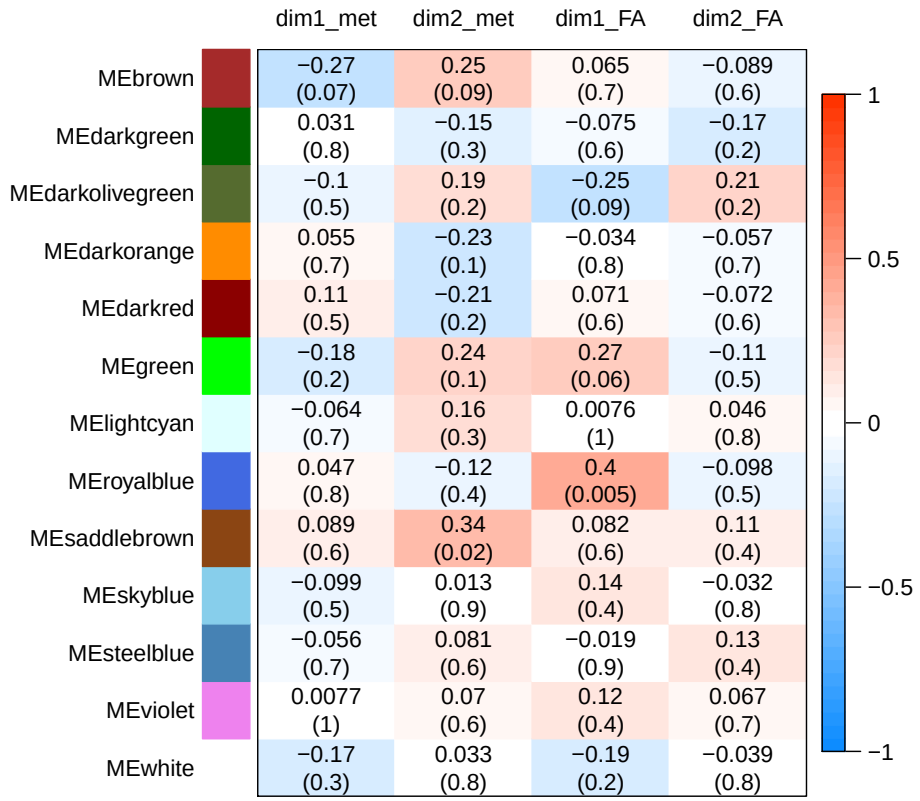

Figure 5.6 – **Heatmap of correlations between molecular modules and profiles of circulating molecules.** Modules of co-expressed probes were obtained from a weighted gene correlation network analysis (WGCNA) from microarray data in the whole blood of 47 growing pigs. The eigengene of each module (ME) was considered as a mathematical representative of the expression levels of the molecular probes within the module. Circulating biochemical molecules were analyzed and the data were summarized by weighted linear correlation using principal component analysis (PCA). The first two dimensions of the PCA were called dim1_met and dim2_met for the metabolites obtained by 1H-NMR high throughput method and dim1_FA and dim2_FA for the fatty acids analyzed by target gas chromatography, respectively.

with dim3, which summarized the circulating concentrations in some amino acids (Gln, Glu and Pro) and hydroxybutyrate. The `saddlebrown` module which tended to be associated with `ADFI` and `%backfat`, was significantly correlated to the second dimension (dim2_met), and also tended to be associated with the fifth dimension (dim5_met) which opposed betaine and TMAO circulating concentrations to inositol concentration. Interestingly, more correlations

were generally observed between dim$i$_met and modules related to %loin: the darkgreen module was highly correlated with dim3_met, and the brown module was correlated with dim3, and to a lesser extent with dim1_met, dim4_met and dim5_met.

Regarding plasma concentrations of FAs, we observed that the module eigengene (ME) of the royalblue module, that was negatively correlated with FCR and positively correlated with ADG, was highly positively correlated to the first dimension (dim1_FA) of the PCA. The darkorange module that was also negatively correlated with FCR, tended to be correlated with the fourth dimension (dim4_FA). The ME of the green also tended to be positively correlated with the first dimension (dim1_FA), whereas ME of the darkolivegreen tended to be negatively correlated.

5.5 Discussion

5.5.1 Analyzing inter-individual variability in feed efficiency

In animal and plant breeding, there has been an increasing interest in intermediate omics traits such as metabolomics and transcriptomics that mediate the effect of genetics on the phenotype of interest [125]. This study confirms that analyzing transcriptome in the whole blood and metabolome in plasma of growing pigs enables to depict the biological molecular pathways involved in various phenotypic traits related to feed efficiency (FE), namely feed conversion ratio (FCR) (i.e., the on-farm measure of feed efficiency) and average daily gain (ADG) (i.e., one of the component of FCR describing growth rate during the test period), and to a lesser extent, body composition (%loin, %backfat). This study is a step ahead for the understanding of the relationships between entities that can act in the inter-individual variability in these traits. Indeed, previous studies have rather addressed differentially-expressed (DE) genes [97, 116], gene set enriched pathways [126] or metabolic signatures [114, 119, 127] between the lowest and the highest FE animals in pigs, cattle or sheep. They have thus compared a gene or a molecule with itself in different conditions such as the response to selection for RFI (a measure of net feed efficiency) or extreme groups based on RFI or FCR. This does not enlighten and explain the interactions between entities in the architecture of the traits, and the behavior of regulatory genes acting in complex traits [128]. Moreover, the aforementioned studies often analyzed only a single type of omics data. Even when they included serum biochemistry in addition to transcriptomics [129] to illustrate consequences of variations in gene expression profiles, they did not intent to depict the correlations between the two levels of life organization

(i.e., gene expression profiles and metabolites). The present study used networks approaches to reveal the main biological processes that are associated with the inter-individual differences in animal traits related to FE. Networks approaches are based on the assumption that the effect of the change in the expression level of one entity can be propagated through the interactions on other entities to orchestrate complex phenotypes [130]. We identified a total of 33 sub-networks (modules) of co-expressed genes in the whole blood across 47 pigs. The network was built with a low threshold set for the minimum number of co-expressed entities in the modules, with the assumption that this may facilitate not only the identification of co-expressed but also of co-regulated genes. Finding one to three clusters of enriched biological processes for the majority of the molecular modules argues for the homogeneity in biological processes shared by the co-expressed genes participating to each module. Furthermore, modules in close-vicinity in the dendrogram (hierarchical clustering), such as the `darkolivegreen`, `lightcyan` and `steelblue` modules or the `royalblue`, `darkred` and `violet` modules, respectively, did not share identical GO terms. This argues for keeping these modules separated rather than merging them. Altogether, the procedure used for network building in this study was then adequate for the identification of co-regulated entities in the different modules.

5.5.2 Enriched pathways in co-expressed genes modules related to variability in feed efficiency

From the 33 modules identified in the whole genes network across the 47 pigs, six modules were significantly associated with `FCR` and two modules tended to be related to `FCR`. In the whole blood of young pigs, another study [116] has previously identified four co-expression modules (minimum of 30 genes per module as threshold, leading to 89 to 786 genes per modules) in the low or high `RFI` groups, and indicated that DE genes overlapped with each of the four differentially expressed modules; however, they found only one module that was significantly correlated to the `RFI` phenotype. Moreover, 34 modules of co-expressed genes were identified [128] from RNA-seq analysis in the liver of low vs high `RFI` cattle (using a threshold of 30 genes as the minimum in each module), out of which four modules showed significant correlations to `RFI`. Importantly, the majority of the modules related to `FCR` in the present study were also related to `ADG` and to `%loin`, whereas none of them were significantly associated with individual feed intake (`ADFI`). This suggests that the main molecular entities in the whole blood explaining the inter-individual variations in FE were involved in the determinism of lean growth potential rather than acting in the regulation of feeding behaviour. The pigs used

in this study originated from a divergent selection for RFI and fed different diets during the test period. Because ADG is an independent variable in the regression that estimates predicted feed intake, RFI and ADG have no correlation. The situation is however quite different for FCR, since ADG is part of the ratio in the calculation. Although Gilbert and colleagues [92] indicated low responses (although statistically significant) to RFI selection on lean meat content across generations of pigs, several studies have consistently found a higher proportion of lean pieces in the most feed efficient pigs as induced by genetic selection [131, 132] or by management strategies [133]. Skeletal muscle is the largest organ in the body and plays important roles in the utilisation and storage of a large proportion of the energy from feed. This likely explains why the molecular modules related to FCR were also partly associated with %loin.

To depict the biological functions of the modules identified in the current study, a functional enrichment analysis was performed within the modules separately. Five of the aforementioned eight modules related to FCR, the violet, lightcyan, darkorange, royalblue and darkolivegreen, were significantly enriched in immune and defense-related processes. In the whole blood, immunity and stress response have been previously identified as biological pathways shared by DE genes between low and high efficient pigs [116]; however, correlation analysis to RFI phenotype rather suggested the importance of a module of co-expressed genes participating to cell adhesion, apoptotic process and immunoglobulin production. In the present study, the importance of immunity and defense-related pathways in the architecture of FCR trait may be over-estimated, since we considered the blood where these processes are specifically enriched. However, previous studies on DE genes in pigs have also reported differences in expression levels of genes involved in defense pathways when examined in different tissues (liver, skeletal muscle, adipose tissue) between divergent lines for RFI [132]. In cattle, [129] also found an enrichment of the transcriptomic networks in the inflammatory response, regulation of monocyte differentiation, proliferation and differentiation of T lymphocytes in the liver from animals with low or high RFI. Since the whole blood reflects the concerted actions of the different tissues, these data support the current findings that immunity and defense-related pathways are important in the determination of feed efficiency. Finally, the recent identification of variations in expression of genes associated with the immune system in milk from low vs high FE dairy sheep [134] , further argues for the informative potential of immune and defense pathways to depict feed efficiency phenotypes of farm animals when analyzed in various biological fluids. Defense mechanisms trigger the use of nutrients for basal metabolism rather than for production performance. This likely explain the importance of defense mechanisms in the determination of feed efficiency, through regulations of ADG and %loin. In support, [135]

shows that high RFI piglets (the less efficient animals) had greater resting energy expenditure and respiratory quotient than low RFI piglets (the most efficient). Among defense mechanisms, we suggested that T cell signaling (royalblue) was negatively related to FCR, whereas B cell activation (lightcyan) and inflammation (darkolivegreen) were positively associated to FCR. This association is likely due to the fact that inflammatory stimulation was associated with a re-orientation of nutrients and alteration of metabolism [135] in growing pigs, thus deteriorating feed efficiency of the animals.
In the present study, other biological processes were also enlightened as important contributors to the inter-individual variability of FCR. Indeed, three modules were composed of co-expressed genes involved in translation (darkred), protein localization (steelblue) or circulatory system development and learning (white). In the whole blood of cattle, a set of genes associated with the metabolism of proteins was also identified as the most enriched pathway of genes differentially inhibited or activated in high-RFI when compared to low-RFI beefs [126]. These pathways could more specifically account for the regulation of lean growth rate, because significant correlations with ADG and %loin were also identified. Finding nitrogen metabolic process (skyblue), protein modification and muscle cell differentiation (brown) as enriched processes in modules related to %loin further support the assumption that whole blood can encompass molecular mechanisms involved in muscle development and metabolism. Finally, two modules, the darkgreen and saddlebrown, were or tended to be related to both ADFI and %backfat, a surrogate of body adiposity. The darkgreen module encompassed genes involved in the response to toxic substances. In accordance, there is generally a marked reduction in voluntary feed intake in disease-challenged pigs [136].

5.5.3 Important genes in molecular networks related to feed efficiency

The main objective of the current study was to enlighten interaction networks related to feed efficiency, rather than focusing on single genes. However, when looking at the hierarchy of the genes in the molecular modules found as underlying FCR, we pointed RPL14 in the darkred module. This gene has been previously suggested by bioinformatics as a hub node gene in regulatory networks [137]. This is an important point to argue for the biological relevance of gene network architecture built herein. Moreover, some of the genes contributing the most to the association between transcriptomics level and animal trait as ranked according to (GS.FCR) values in each module, have been previously identified as genes with fold changes

in their expression level greater than $|2|$ between groups of low and high (RFI) pigs [97]. There were SLPI (violet module), EIF1B (darkred module) and HTRA1 (ligthcyan module). As expected, these genes are participating to different biological processes listed as specifically enriched in their parent modules, such as immune response by protecting epithelial surfaces (SLPI), regulation of cell growth (HTRA1) and translational initiation (EIF1B). Of note, HTRA1 encodes a secreted enzyme that may regulate the availability of insulin-like growth factors (IGFs), and correlated responses of IGF-I to RFI have been observed in pigs [138]. Altogether, these genes are likely biologically important in the variability of FCR.

The royalblue module is of upmost interest since it was associated with FCR and with profiles of circulating fatty acids (see next section). Therefore, the molecular functions of its top-ranked genes deserve deeper investigations in the Human Gene Database. The LPAR3 (Lysophosphatidic Acid Receptor 3) and CCR7 (C-C Motif Chemokine Receptor 7) genes are members of the G protein-coupled receptor family (GPCR). Especially, the protein encoded by CCR7 is known to activate B and T lymphocytes and to control the polarization of T cells in chronic inflammation. Another gene involved in T-cell signaling was SKAP1 (Src Kinase Associated Phosphoprotein 1) that is required for an optimal conjugation between T cells and antigen-presenting cells. Although EPHB6 (EPH Receptor B6) codes for a protein that mainly influences cell adhesion and migration and regulates cell developmental process rather than immunity, one of its related pathways is GPCR signaling. Interestingly, the GPCR pathway has been also identified as a putative candidate for RFI difference in pigs by genome-wide association studies [139]. Another member of the royalblue module is NPY, a gene coding for the neuropeptide Y that influences many physiological processes including stress response, food intake, energy balance and circadian rhythms. In accordance, hypothalamic genes expression including NPY plays a potential role in feed efficiency variation in different farm species [140]. The neuropeptide Y also functions through GPRC. Finally, among the top genes in the white module, HTR7 encodes the serotonin receptor which belongs to the GPCR family and is regulating several behaviours of animals.

5.5.4 Relationships between transcriptomic and metabolic levels in the definition of feed efficiency or related traits

Combined phenotype-metabolome-genome analysis by inferring gene networks based on partial correlation and information theory approaches has been valuable to confirm cellular maintenance processes as major contributors to genetic variability in bovine feed efficiency

[141]. Therefore, the present study also addressed the interactions between two levels of organization in the circulating entities, i.e., transcriptome and metabolome, and their relations to productive traits in the pigs. Variations at the metabolic level were first summarized by linearly transforming the data into few new coordinates that explained most of the total variance, thanks to principal component analysis (PCA). Correlating the PCA coordinates to the eigengenes of the 33 modules allows to determine whether pigs with the same profiles in circulating fatty acids (FA) and lipoproteins, amino acids (AA) or energy-related metabolites (glucose, lactate, betaine, etc.) shared similarities in groups of co-expressed genes. However, few significant correlations were identified between the two organization levels. The eigengene of the `royalblue` module was highly correlated to the PCA coordinate that associated circulating concentrations in saturated FA and in polyunsaturated FA (omega-6 FA, and to a lesser extent, omega-3 FA families). In other words, the greater expression levels of genes involved in T signaling, cell adhesion and cell developmental process in the whole blood, were associated with a higher proportion of saturated FA and a lower proportion of PUFAs in plasma, and altogether, these changes accounted for a better feed efficiency (lower `FCR`) and higher `ADG`). An important regulatory element underlying this association might be the expression level of `LPAR3`, a gene that is involved in phospholipid binding. Similarly, the eigengene of the `darkorange` module which was composed of co-expressed genes related to immune process, tended to be correlated with the circulating concentrations of omega-3 FA, and these processes simultaneously accounted to `FCR`. Among the top-ranked genes in this module, `BMP6` encodes a secreted ligand of the transforming growth factor (TGF-beta) superfamily of proteins that regulate a wide range of biological processes including fat cell development, and `TAOK3` encodes a serine/threonine protein kinase that activates the p38/MAPK14 stress-activated MAPK cascade, a pathway regulating also adipose cell development and metabolism. Thus, whereas the gene network approach did not identify any enriched pathways related to lipid metabolism in the whole blood transcriptome, the combination between transcriptomics and metabolic data suggests that fatty acid metabolism in different tissues can be related to `FCR` and further influenced/be influenced by interconnected molecular pathways of genes related to immunity. In support, muscle of high-FE pigs exhibited lower proportion of saturated FA and an enhanced proportion of polyunsaturated FA when compared with low-FE pigs [131], and co-expression analysis in the liver has revealed altered lipid metabolism between high and low feed efficient steers [129]. Relationships between FA metabolism and immunity have been also described in the literature, showing that omega-3 (n-3) PUFA can suppress T cell antigen presentation, activation, proliferation and cytokine expression [142]. In addition, high fat western diet promotes inflammation

and modifies immunity [143].

Among significant associations identified between modules of co-expressed genes in the whole blood and profiles in circulating metabolites, the darkgreen module composed of genes involved in the responses to lipopolysaccharide (LPS) and toxic substances and related to ADFI, was associated with the circulating concentrations of beta-hydroxybutyrate and of Pro, Glu and Gln amino acids. Beta-hydroxybutyrate is a ketone body whose concentration is rising up after fasting and in situation of energy deficit, illustrating its dependence to the regulation of feed intake. Sensing of AA can also act on the hypothalamic control of food intake [144]. Networks of co-expressed genes related to %loin (such as brown and green) also displayed significant correlations with the metabolome profile, which suggests strong relationships between AA metabolism and the molecular regulation of muscle growth in the pigs. In support, protein (amino acids) metabolism is essential for optimizing efficiency of nutrient absorption and metabolism and to enhance growth performance. Relationships between transcriptomics and metabolomic data to depict the biological processes underlying complex production traits like FCR, ADG, ADFI, have been identified herein by statistical analyses (multivariate-based procedures for data concatenation) and then, scrutinized with the functional annotation tools DAVID (pathway-based integration techniques) and expert knowledge about the potential roles of specific entities. However, when transcriptomic and metabolomic data are integrated, there is no direct association between metabolite and transcript. Although commercially available tools have been developed to visualize ranked pathways among molecules, there are many biases when treating genes and metabolites as equivalent entities [27]. To explore the causality within the interconnected entities at the different levels of cell organization, it seems necessary to use the graph theories. First, the feed efficiency networks identified herein could be compared in their topologies (direct interactions, connectivity degree per gene, etc.) with random networks [141]. Second, knowledge graphs can be generated thanks to web semantic-dedicated queries to identify paths composed of chains of relationships. Path lengths between entities (pairs of co-expressed genes and small molecules), traversed properties (edges) and encountered biochemical reactions could be then analyzed. However, the mapping between different identifiers of genes/metabolites in naming systems is still a problem to be overhelmed in this process [27].

In conclusion, the inter-individual differences in feed conversion ratio (FCR, i.e., the on-farm measure of feed efficiency), were inferred to be mainly due to variation of co-expressed genes participating to immunity, defense mechanisms, inflammatory response, cell developmental process, translation and protein localization. These variations induced changes in the capac-

ity of amino acids usage and lipid (fatty acids) metabolism between pigs. Among the component traits of FCR, these processes accounted likely more in the variation of growth rate than in the regulation of feed intake. However, few genes in the gene networks (e.g., NPY) are suggested for their roles in regulating feeding behaviour. Analyzing the gene network also allowed to propose integrative regulatory mechanisms such as G protein-coupled receptors (GPCR). Relationships were indicated between T cell receptor signaling, cell development process and circulating concentrations of omega-3 fatty acids in plasma, which both underlined inter-individual variability in feed efficiency. This suggests that nutritional recommendations for growing pigs should consider the lipid fraction of diets to improve health and production traits in synergy.

Declarations

Ethics approval and consent to participate

This study was based on previous published studies. The original publications have included a statement on ethics approval to use animals into genetics and feeding experiments. In these publications, the care and use of pigs were performed in compliance with the European Union legislation (directive 2010/63/EU). All animals were reared in compliance with the French national regulations on ethics in animal experimentation (Décret n° 2013-118 du 1er février 2013 relatif à la protection des animaux utilisés à des fins scientifiques), and killed according to procedures approved by the French veterinary services. The protocols were approved by an Ethics Committee in Animal Experiment (Comité Rennais d'Ethique en matière d'Expérimentation Animale, agreement N°07–2012; Ministère de l'Enseignement, de la Recherche et de l'Innovation) from which INRAE is affiliated. All methods have been reported in accordance with ARRIVE guidelines (Animal research: Reporting of in vivo Experiments). The present study follows the internal charts on ethics and animal experimentation at INRAE since we only reused data and did not include additional living animals.

Availability of data and materials

Transcriptomic dataset was obtained from NCBI's Gene Expression Omnibus (GEO) Subserie accession number GSE70838 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE70838). The metabolomic dataset was retrieved in Jegou et al. [119]. Phenotypic traits and fatty acid composition in plasma are are deposited in the publicly available repository at https://entrepot.

recherche.data.gouv.fr/dataset.xhtml?persistentId=doi:10.57745/TM2ANC.

The adjacency matrix is available at https://data-access.cesgo.org/index.php/s/YPz0J2ItxIEuN5M.

We provide R notebooks detailing the analysis https://github.com/cjuigne/multiomics_and_feed_efficiency.

CHAPTER 6

A GRAPH-BASED APPROACH TO IDENTIFY COMPLEX CONNECTIONS IN HETEROGENEOUS BIOLOGICAL NETWORKS

This chapter is an article being prepared for submission as an original research article to *Oxford Bioinformatics*.

➡ **Juigné C**, Becker E, Carpentier O, Gondret F, Dameron O. "A graph-based approach to identify complex connections in heterogeneous biological networks". In preparation for *Bioinformatics*.

Abstract

Multimodal analysis of biological systems using metabolomics and transcriptomics has gained significant interest. However, integrating diverse and massive heterogeneous data, along with complex reasoning, poses challenges. In this study, we propose a method to integrate multimodal -omics data using Semantic Web technologies. Our analysis method based on Reactome BioPAX export identifies knowledge-based chains of relationships between statistically related nodes in biological datasets, facilitating explainability and suggesting modulatory biological actors. We applied our method to a set of co-expressed genes related to feed efficiency as a use case. The results show that a large percentage of proteins (89%) and small molecules (70%) can be linked to UniProt IDs and ChEBI identifiers, respectively. The paths between genes in the co-expressed gene networks are shorter compared to randomly produced networks of the same size. This work offers new possibilities for integrating multiple -omics data types and discovering networks between molecules and potential upstream regulators. These networks provide insights into the underlying biological processes of co-expressed genes.

6.1 Introduction

High-throughput -omics techniques like transcriptomics (genes expression levels), proteomics (proteins) and metabolomics (metabolites and small molecules), as well as target analyses for specific molecules generate large amounts of multimodal data. Typically, each modality can be statistically analyzed to produce lists of differentially-expressed molecules between experimental conditions.

We hypothesize that considering the different levels of -omics as a whole will support a broader and finer understanding of the processes governing biological systems. Indeed, to provide a holistic view of the tissue or cell behaviors, it is valuable to obtain an extensive description of where and how all types of molecules participate and interact with each other within and between biological pathways. This systemic representation provides a better knowledge of the cascade of events, the upstream regulators of specific pathways (series of interactions), and the cross-talks between pathways.

Among the approaches dedicated to the joint analysis of heterogeneous -omics data, matrix factorization has proven to be particularly efficient [37]. Each -omics dataset is represented as a very large matrix that will be further decomposed into a product of two matrices, with the constraint that one of the matrices in the product is common to the different datasets [145]. This dimensionality reduction identifies the most important joint traits in the heterogeneous data. However, these statistical associations often lack biological interpretability and fail to account for the known physical relationships between molecules, such as activation, inhibition, interaction, control, or participation in common reactions.

To address this limitation, we introduce a methodology that maps high-throughput transcriptomics and metabolomics data onto a graph representing knowledge of cell metabolism, including interactions and their regulation. By leveraging the connections between small molecules and proteins, this graph serves as a suitable framework for integrating metabolomic and transcriptomic data with biological process knowledge. The mapping of different -omics levels onto this graph represents an initial step in identifying cascades of reactions associated with specific phenotypes based on a list of multimodal molecules (proteins and metabolites).

6.2 Background

6.2.1 Databases and ontologies in biology

With the expansion of high-throughput biological analysis technologies, significant efforts have been directed towards organizing, storing and sharing data. As a result, there are many different databases and formats that facilitate access and use of these massively produced data [146]. Many of them exploit the semantic web: a large fraction have a SPARQL endpoint [147] and there are numerous ontologies to structure knowledge. Noteworthy examples include UniProt for proteins, ChEBI for small molecules, MI for annotating molecular interactions, and Gene Ontology (GO) for describing biological components and functions (Whetzel et al., 2013). This rich landscape of resources opens up possibilities for integrating multiple knowledge bases and establishing cross-references with external sources.

6.2.1.1 The UniProt database

UniProt is a specialised knowledgebase for proteins maintained by the UniProt consortium [148]. It is known for its excellent quality, thanks to a high level of manual curation by an expert biocuration team. It combines data from several databases and contains metadata and cross-references to other databases and knowledgebases such as publications, sequences, functions, etc. Each protein (i.e., gene product) has a unique and universal identifier, which means that if a single gene encodes more than one form of the protein or if genetic changes occurs, each protein has a specific identifier. UniProt is fully accessible and queryable *via* a SPARQL endpoint.

6.2.1.2 The ChEBI Ontology

ChEBI is a large database dedicated to chemical compounds and an ontology that aims to organize these chemical compounds by biological properties: role, nature... [18]. ChEBI is manually annotated from the literature and each chemical compound has a unique identifier. Cross-references to other resources (Rhea, NMRShiftDB, BRENDA) are also possible and an OWL export of the database is provided [149].

6.2.1.3 The BioPAX ontology

BioPAX is a well established formalism to represent biological pathways at the molecular and cellular levels including interactions [96]. BioPAX is based on Semantic Web technologies, with RDF facilitating integration, SPARQL facilitating querying and OWL facilitating knowledge-based reasoning. All the major pathways databases are available in BioPAX and can be mapped with other resources such as ChEBI [18], UniProt [150], but also Gene Ontology [151] for genes, as well as ontologies for phenotypes or diseases.

In the BioPAX ontology, under the root class named `Entity`, the four top level classes are `Pathway`, `Interaction`, `PhysicalEntity` and `Gene`. Interactions capture biological relationships involving two or more entities, including molecular interactions, controls, and conversions. Physical entities encompass various components such as small molecules, proteins, DNA, RNA, and complexes.

BioPAX introduces the class of `EntityReference` as "*a grouping of several physical entities across different contexts and molecular states, that share common physical properties and often named and treated as a single entity with multiple states by biologists*". This means that there is a corresponding entity reference node for proteins, small molecules, DNA, DNA regions, RNA and RNA regions, where all the non-changing aspects of the entity are stored. The BioPAX specification mentions that "*there should only be one `EntityReference` defined per UniProt ID*". The physical entities are linked to these entity references by the property `entityReference`, and one `EntityReference` can relate to several `PhysicalEntities`.

In the BioPAX ontology, there are different utility classes intended to annotate the `Entity`. Among them, we can find the `Xref` class defined as "*A reference from an instance of a class in this ontology to an object in an external resource.*" in the BioPAX documentation. This allows references to external databases or ontologies, for example by providing a UniProt identifier to a protein or a Gene Ontology term to a physical entity. Each node `Xref` has the following properties: `db` which reference the external database, `dbVersion` to explicitly provide the database version and `id` which links the physical entity in BioPAX to the external reference. One `EntityReference` can be annotated with various references from external sources via these `Xref` nodes.

6.2.1.4 The Reactome knwoledgebase

Reactome [1] is a free, open-source, curated and peer-reviewed multi-species pathway knowledgebase [16]. It is widely used in genome analysis, modeling, systems biology, clinical research and education. Biological pathways can be explored to shed light on inter-connected proteins. Reactome representation is centered on biochemical reactions that can consume and produce proteins, small molecules, or complexes composed of several proteins or small molecules. Additionally, Reactome captures the regulatory mechanisms exerted by entities such as proteins and complexes on these biochemical reactions. Notably, Reactome provides an export feature that allows users to obtain the database in the BioPAX format, ensuring compatibility and interoperability with other systems.

The present study is based on a revised version of Reactome BioPAX files. Indeed, we identified that Reactome's BioPAX export did not comply with the BioPAX standard [152]. This nonconformity had a significant impact on the topology of the interaction graph, artificially increasing the path length between nodes and changing the degree of the `Complex` nodes. Moreover, it obscured implicit redundancy between Complex nodes. By addressing and rectifying these non-compliant aspects (Juigne et al., 2023), we are able to propose a revised version of Reactome's BioPAX files, which now align with the BioPAX standard.

6.2.2 Weighted co-expressed gene networks coupled to metabolic profiles as a use-case

In a previous study, we investigated the biological processes associated with variations in feed efficiency, which reflects the utilization of feed nutrients to support growth rate, in pigs [153]. This analysis involved the examination of transcriptomics data (obtained through microarray) [97] and metabolomics data (utilizing 1H-NMR) [119] from the blood samples of 47 pigs. These data are briefly summarized hereafter, as they serve as a use-case of the methodology presented. For a complete study description, see [153].

6.2.2.1 Weighted co-expressed gene networks involved in feed efficiency

We applied weighted gene co-expression network analysis (WGCNA) to identify 33 weighted co-expressed gene modules. These modules represent clusters of genes that exhibit similar expression patterns across different samples [56]. Out of these modules, we found that the eigen-

1. https://reactome.org/

genes of 8 modules, which summarize the expression profiles within each module, were significantly correlated with feed efficiency traits in the animals [153]. However, the challenge lies in elucidating the underlying biological mechanisms within each module to determine whether the modules consist of co-regulated genes or if the observed correlations are merely coincidental. This distinction is crucial in understanding the true relationships and unraveling the biological significance of these co-expressed gene networks.

6.2.2.2 Integration of transcriptomic and metabolic -omics levels on the same Reactome graph to explore connections

In this study, our proposal involves mapping various organizational levels, specifically molecules from ChEBI and proteins from UniProt, onto the biological reactions extracted from Reactome. By combining these different -omics levels into a unified graph, we take an initial step towards exploring and analyzing the pathways. We explore how this integrated graph can reveal the biological connections and interactions from weighed co-expressed genes modules statistical analysis.

We focused on the latest version of the Reactome pathways database *Homo Sapiens* (version 84 (2023-04)), which is also appropriate to study pigs (*Sus scrofa*). Models of correspondence between human and pig genomes in both material and formal types have been recently reviewed in [154].

6.3 Methods

Overview

We developed federated and semantically-rich SPARQL queries to identify specific physical entities within the BioPAX export of Reactome (detailed in sections 6.3.1 and 6.3.2). These queries are based on a large variety of classes and properties specific to the BioPAX ontology and take advantage of the capability to simultaneously query multiple databases through SPARQL endpoints. The whole integration schema is represented in Figure 6.1. Subsequently, we loaded the integrated graph into Neo4j and performed traversal Cypher queries (section 6.3.3) to characterize the paths between molecules of interest.

To ensure reproducibility, we provide a Jupyter notebook detailing the analysis [2].

2. https://github.com/cjuigne/data_integration_biopax

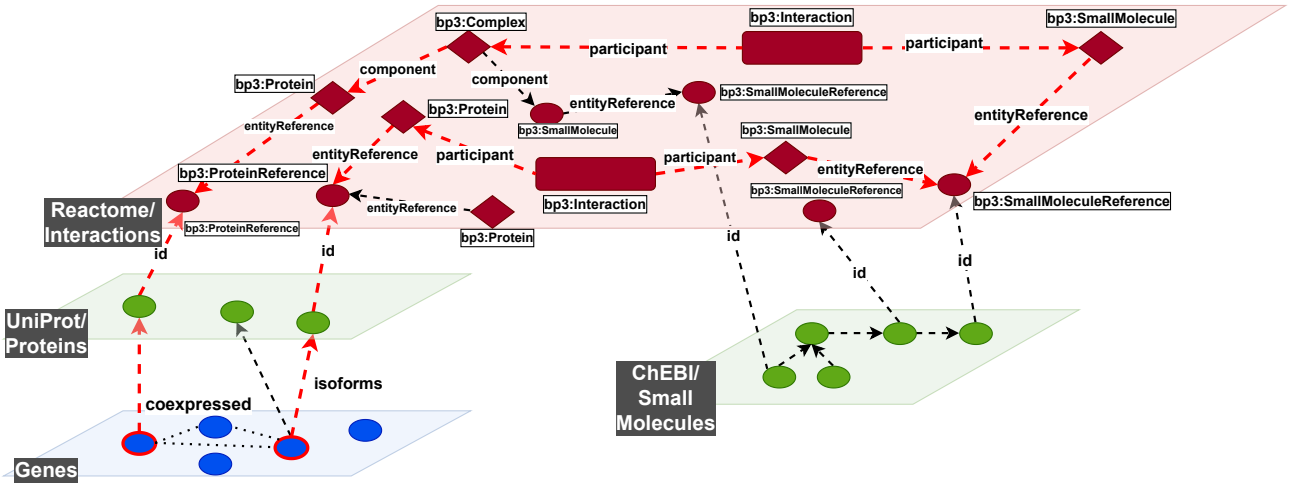

Figure 6.1 – Integration schema.
Each layer represents a type of data from a data source linked to the other layers by different properties. Traversing the relations within and between layers provides some candidate biological interpretation of the processes involving co-expressed genes (e.g. in red a simplified path between two of the three co-expressed genes, see Fig 6.2 for a realistic example). Note that some layers (e.g. ChEBI) provide relations among some of their nodes, so the integration of the different layers results in richer graph traversal capabilities by connecting nodes that were previously disconnected (e.g. in the Reactome layer) or by providing alternative paths.

6.3.1 Retrieving `Gene` and `Protein` in the BioPAX export of Reactome using federated SPARQL queries

As introduced in section 6.2.1.3, `Proteins` are linked to a `ProteinReference` *via* their `entityReference` property. These `ProteinReferences` are themselves linked to a cross-reference `UnificationXref` providing their UniProt identifier (Figure 6.2). We use a 3-steps federated SPARQL query that takes a list of HGNC IDs as input:

— First, it queries the SPARQL endpoint of the UniProt database to get all UniProt IDs related to a HGNC ID.

— Second, it looks for the corresponding `ProteinReferences` in the BioPAX file.

— Finally, it identifies all the `Proteins` associated to each `ProteinReference`.

This process ensures the establishment of connections between HGNC IDs, UniProt IDs, `ProteinReference` objects, and associated `Protein` entities.

6.3.2 Retrieving `SmallMolecules` in the BioPAX export of Reactome using federated SPARQL queries

As introduced in section 6.2.1.3, `SmallMolecules` are linked to a `SmallMoleculeReference` *via* their `entityReference` property. In the Reactome dataset, `SmallMoleculeReferences` are annotated with ChEBI identifiers via `UnificationXref` nodes (Figure 6.2).

Starting from a list of metabolites names, we used a SPARQL query to retrieve of the molecule and its possible enantiomer, as well as their respective descendants.

Then, with the next SPARQL query, we identified the corresponding entities in Reactome. Precisely, for each molecule, we looked for the `SmallMoleculeReferences` that are annotated with the corresponding ChEBI IDs; then, we identified all the `SmallMolecules` associated to these entity references.

This strategy wasn't conclusive, as it returned to many ChEBI entities candidates. We thus selected manually the identifiers that seemed to be the most appropriate ones and we are currently working on fixing that problem so that the federated query works.

6.3.3 Computing paths between nodes of interest in Neo4j using Cypher queries

After the reference identifiers of the molecules of interest have been unified by federated SPARQL queries and the URI of the nodes of interest have been extracted, we used the NeoSemantics plugin to load the integrated graph in BioPAX format into a Neo4J database. This allows to combine the semantic richness of the BioPAX ontology, and at the same time, to benefit from graph traversal capabilities of the Cypher query language.

We designed graph traversal queries to characterize the paths between pairs of target nodes belonging to the same module of co-expressed genes. We searched for the shortest paths between two proteins of interest (from the same module of co-expressed genes) to focus on the metabolic relationships that could bring a different perspective to the co-expression.

To retrieve biologically-compatible paths, our queries specify the combinations of properties that can be traversed and those that should not be. Indeed, paths are only allowed to pass through `Complex` nodes *via* the `component` property, through `Interaction` nodes (and its sub-classes) *via* the `participant` property (and its sub-properties), and through `EntityReference` (and its sub-classes) via the `entityReference` property. Traversal

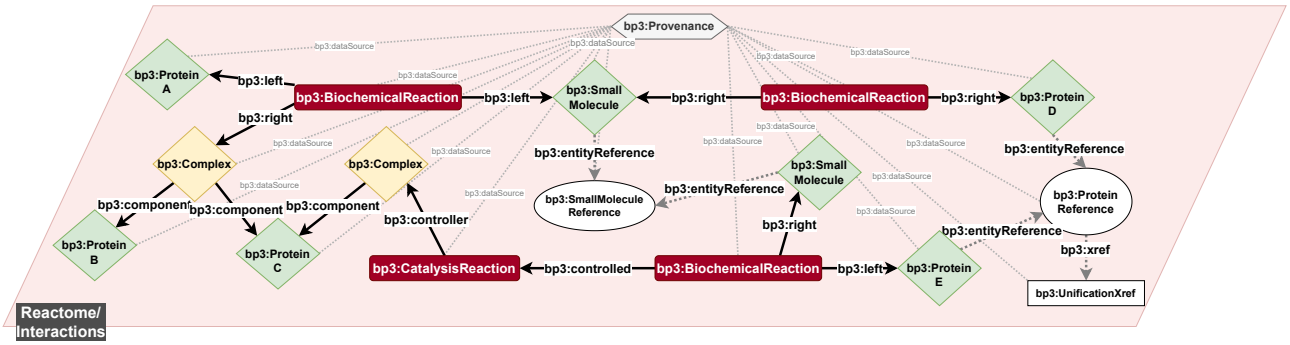

Figure 6.2 – Examples of allowed and disallowed path patterns.
The shortest path between *Protein B* and *Protein C* is length two and consists of two `bp3:component` edges. The shortest path between *Protein C* and *Protein E* traverses two *Interaction*s thanks to the subproperties of `bp3:participant`. The shortest path between *Protein A* and *Protein D* contains a *SmallMolecule*, this type of path demonstrates the usefulness of using a metabolic network. Protein D and E have the same UniProt ID provided by their *ProteinReference*. In Reactome, all nodes are connected to the *Provenance* node which indicates (through its properties) that all data comes from the Reactome database. There are other properties and classes that connect many nodes, thus, we avoid this type of path.

through other properties is forbidden because it might lead to irrelevant paths. For example, all nodes in Reactome are connected to the `Provenance` node which indicates through its properties that data were extracted from the Reactome base. Thus, all couple of proteins might be at distance 2 if the `datasource` property used to reach the `Provenance` node is allowed.

For each couple of protein of interest, we retrieve the number of shortest paths (there might be several ones), the shortest path length, the properties traversed, the number and identifiers of biochemical reactions traversed and the number and identifiers of small molecules traversed.

6.3.4 Biochemical reaction cascades and their regulation in modules of co-expressed genes and comparison with randomizations

To capture properties specific to co-expressed gene modules from experimental transcriptomics data, we compare the shortest paths within each WGCNA module to the shortest paths calculated inside 500 randomly generated modules of the same size. For each randomization, we randomly pick the same number of `ProteinReference` nodes in the entire graph, and compute the shortest paths between all the pairs of these `ProteinReference`. Then, we compare the shortest paths distributions inside one co-expressed gene module from experimental transcriptomics data to the shortest paths distribution inside the 500 randomly-generated

modules using a non-parametric Mann-Whithney test. We also compare the number of biochemical reactions that have to be traversed using a non-parametric Mann-Whithney test. As there might be several shortest paths connecting two proteins, we study:

— the number of biochemical reactions on the shortest paths;

— the proportion of paths containing at least one small molecule.

All statistical tests were performed using R software environment.

6.4 Results

6.4.1 Proteins retrieved by their UniProt ID in Reactome

Reactome v84 is composed of a total of $31,332$ `Proteins` and $11,672$ `ProteinReferences`.

Among the $11,672$ `ProteinReferences`, $11,299$ are linked to a unique UniProt ID (*i.e.* 97%). In accordance with the BioPAX specification (6.2.1.3), several `Proteins` can point to the same `ProteinReference`. Up to $1,328$ distinct `Proteins` from Reactome are linked to the most connected `ProteinReference`, which unsurprizingly the cellular tumor antigen p53 (UniProt ID: P04637).

When focusing on `Proteins`, we observed that $27,453$ of the $31,332$ (*i.e.* 87%) have a UniProt ID provided by their `ProteinReferences`. Some `Proteins` without a UniProt ID associated to their `ProteinReference` (in the remaining 13%), can be associated with a UniProt ID when their `name` property directly contains a UniProt ID. This happens for 287 `Proteins`, that all are minor isoforms. Overall, we are able to associate $27,740$ `Proteins` (88%) from Reactome to their UniProt ID in Reactome.

Among the 8 modules of co-expressed genes selected for their associations with the animal phenotype (i.e., feed efficiency), we retrieved from 56 to 100% of the UniProt IDs in Reactome (Table 6.1).

6.4.2 SmallMolecule retrieved by their ChEBI ID in Reactome

Reactome v84 contains $5,083$ `SmallMolecule` entities and $2,903$ `SmallMoleculeReference` nodes. $1,946$ of these $2,903$ `SmallMoleculeReferences` are linked to a unique ChEBI ID (*i.e.* 67%). There are from 1 to 28 `SmallMolecules` pointing to a `SmallMoleculeReference`. The most

Table 6.1 – Number of UniPort IDs corresponding to co-expressed genes in the eight modules of interest and number of corresponding *ProteinReference*s and their *Protein* instances retrieved in Reactome.

Modules	**white**	**darkorange**	**royalblue**	**violet**	**darkred**	**darkolivegreen**	**steelblue**	**lightcyan**
Corresp. Proteins in UniProt	23	14	71	12	43	9	10	95
Corresp. Protein References in Reactome	13	10	49	10	28	9	7	66
Protein instances in Reactome (BioPAX)	61	14	200	39	33	22	18	127

connected `SmallMoleculeReference` is H^+ (CHEBI ID:15378), with 28 connections. When focusing on `SmallMolecules`, we observe that $3,562$ of the $5,083$ (*i.e.* 70%) have a ChEBI ID provided by a `SmallMoleculeReference`.

In our use-case metabolomic dataset from whole blood of pigs with differences in feed efficiency, 50 metabolites were investigated [119]. From the 50 metabolites names of interest, we retrieved 49 ChEBI IDs corresponding to 165 ChEBI elements when we considered sub-elements and enantiomers (from 1 to 30 IDs per molecule of interest). We found 27 of these IDs in BioPAX (from 1 to 4 `SmallMoleculeReference` per 20 molecules of interest). This is a first step to study whether or not these small molecules are close to the targeted *Protein* nodes and/or on the paths linking these proteins (to be continued).

6.4.3 Graph traversal and paths connecting molecules of interest

We compared the shortest paths connecting members of a common co-expression module, to the shortest paths between the members of randomly-generated modules of the same size. For the largest module (named `lightcyan` in Table 1), computations failed due to its large size. Among the other seven co-expression modules studied, we observed two distinct behaviors.

For some of them such as module `darkred`, the WGCNA module exhibited shorter paths than randomly-generated modules of the same size. Fig. 6.3 (middle) shows a significant shift of the distribution of the shortest paths lengths between the co-expression module and the 500 randomizations (p-value $< 2,2e^{-16}$, average length in module 7.2 *vs* 9.6 in randomizations). We also observed that shortest paths inside WGCNA module `darkred` traversed significantly fewer biochemical reactions than inside randomly-generated modules (p-value $= 3.7e{-}9$, average biochemical reactions within shortest path in module 1.5 *vs* 2.3 in randomizations). Similar features are also observed for WGCNA module `royalblue` presented in Fig. 6.3 (left).

For other co-expression modules such as WGCNA module `white`, these properties did

not differ from those found in randomly-generated modules (statistical tests not significant). This is illustrated in Fig. 6.3 (right). The modules corresponding to this behavior are the smallest ones, with less than 15 corresponding protein references in Reactome (modules `white`, `darkorange`, `violet`, `darkolivegreen`, `steelblue`).

We investigated whether the shortest paths connecting proteins within co-expression modules exhibit specific topologies compared to randomly identified paths. Our focus was on the number of biochemical reactions occurring along these paths and the count of metabolites constituting these paths. We calculated the percentage of paths that did not traversed any `BiochemicalReaction` node, passed through exactly one `BiochemicalReaction` node, and traversed at least one `BiochemicalReaction` node. Additionally, we determined the percentage of paths containing at least one `SmallMolecule` node. The overall results are presented in Table 6.2.

In random modules, the proportion of shortest paths passing through at least one biochemical reaction varied widely, especially in smaller modules, ranging from an average of 28.2% for randomly generated modules of size 7 to 75.9% for those with a size of 9. For the two sets of largest modules, the average percentage of paths involving at least one biochemical reaction spanned from 33.4% for random modules of size 28 to 38.32% for random modules of size 49. Remarkably, within the two largest co-expression modules (`royalblue` and `darkred`) paths between proteins pairs were more likely to involve biochemical reactions than random behavior. Specifically, 41.5% of the paths in the module `royalblue` involved biochemical reactions, in contrast to 38.32% of shortest paths in randomly generated modules. In the `darkred` module, 76.44% of shortest paths contained biochemical reactions, compared to the 33.4% observed in random modules.In random modules, the percentage of shortest paths passing through a metabolite ranged from 10.75% to 29.28%. This percentage exhibited greater variability in smaller modules, while in the randomizations of the size of the two largest modules (`royalblue` and `darkred`) it was more consistent: 24.3% for `royalblue` and 20.6% for `darkred`. For these significant gene co-expression modules, the percentage of paths involving a `SmallMolecule` node exceeded that of random modules: 41.2% of shortest paths between protein pairs in `royalblue` involved a small molecule, and up to 82% of shortest paths between protein pairs in `darkred`. Similar trends were observed in most smaller co-expression modules, except for `darkorange` and `steelblue`.

Table 6.2 – Analysis of the topology of the shortest paths in co-expression modules compared with randomly generated modules of the same size. PR : `ProteinReference`, SM : `SmallMolecule`, BR : `BiochemicalReaction`.

	white (PR =13)	**darkorange (PR=10)**	**royalblue (PR=49)**	**violet (PR=10)**	**darkred (PR=28)**	**darkolivegreen (PR=9)**	**steelblue (PR=7)**
Paths without BR	Co-expression modules						
	23.81%	0%	58.51%	0.79%	23.56%	0.68%	10.72%
	Randomly generated modules						
	54.15%	46.98%	61.68%	52.96%	66.60%	24.13%	71.77%
Paths with BR=1	Co-expression modules						
	20.24%	0.35%	6.50%	0.63%	17.48%	1.08%	88.63%
	Randomly generated modules						
	24.03%	19.10%	18.35%	26.92%	16.05%	46.37%	13.12%
Paths with BR≥1	Co-expression modules						
	76.19%	100%	41.49%	99.21%	76.44%	99.32%	89.28%
	Randomly generated modules						
	45.85%	53.02%	38.32%	47.04%	33.40%	75.87%	28.23%
Paths with SM≥1	Co-expression modules						
	98.68%	1.04%	41.18%	99.75%	82.84%	99.76%	1.18%
	Randomly generated modules						
	19.82%	29.28%	24.97%	29.97%	20.65%	18.23%	10.75%

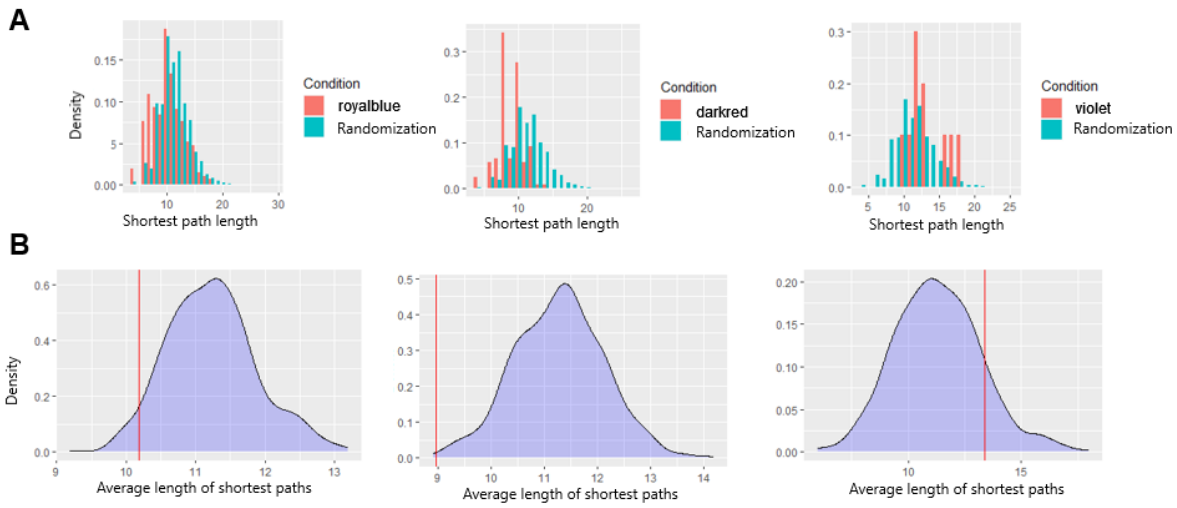

Figure 6.3 – **A.** Distribution of the shortest path length in co-expressed gene modules `royalblue`, `darkred` and `violet`, compared with the average shortest path length of 500 random modules, between pairs of `ProteinReference` nodes within a module.
B. Distribution of the average length of the shortest paths in random modules. The red line represents the average length of the shortest paths between pairs of `ProteinReference` in the associated WGCNA module.

6.5 Discussion

Mapping weighed co-expressed gene modules in a whole metabolic network should provide valuable biological information supporting statistical links, and thereby extending co-expression to co-regulation of genes behind animal phenotypes.

We analyzed 7 modules that were highly correlated to the phenotype of interest (feed efficiency defined as the ability of an animal to use nutrients to support growth rate) in our companion experimental study [153].

In this study, we compared the distance between pairs of nodes from 7 statistically co-expressed modules to the distance between pairs of nodes in randomly-generated modules of the same size.

For 2 of the 7 modules, the distribution of distances between pairs shows a particular behavior in co-expressed modules. Inside these modules, the average distance was lower than for the randomly generated modules. Interestingly these two modules are the largest ones (49 and 28 `ProteinReferences`). Their sizes contribute to reaching statistical significance.

On the opposite, in the other 5 modules analyzed in the use-case, the distances between co-expressed genes were not different from the distance between randomly-generated modules. We note that non-significant modules were the smallest ones, with less than 15 corresponding protein references. Thus, lack of significance might be due to small size, but can also reflect that the relationship between the genes was co-expression not driven by a common co-regulation.

These graph traversals give us a valuable indications on the underlying metabolic relationships existing between set of co-expressed genes, thanks to gathering together different levels of cell organization. However, the method could still be improved, by questioning the specificity of the nodes. Indeed, in the interaction graph, there are hubs that are connected to a large number of the nodes. Some of them can be considered as too generic, such as water or H+ because there are used by the majority of biochemical reactions. This creates artificially short paths. Possible solutions could be to declare some hubs in a black list or, in a softer way, to add weights on the graph. Additionally, in biological networks, the shortest path between two molecules is not always the most effective path. Flux distribution of within the different branches of the network occurs in many biological situations. Quantification of the energy efficiency of different nutritional scenarios have been proposed by others [155]. In this study, we computed shortest path but did not prove that this strategy was the most appropriate. In the future, other paths could be considered based on specificity or quality indices.

Other types of biological networks exist based on specific data type. Protein-Protein Inter-

actions (PPI) networks, such as Intact [156], could also be used to assess the proximity between proteins. However, our study proves that using metabolic network that includes several data modalities, is relevant: among the identified paths, we retrieved some that contained *Small-Molecule*. In the future, we will investigate the presence of the small molecules of interest on the shortest paths identified, and seek to link the proteins to these small molecules. As a perspective, graph-based approach may be used to identify relevant biochemical paths between unimodal or multimodal data. This will help to focus on the biochemical interactions encountered on these interesting paths with the aim of understanding the link between our modules and metabolism.

In the current study, the Semantic Web provides the framework for integrating transcriptomic and metabolomic experimental data with the knowledge on biochemical interactions, proteins and molecules. The BioPAX ontology allows to represent finely the biological interactions at both cellular and molecular levels, thanks to a well designed data schema. However, BioPAX remains a fairly complex format description and raised several challenges.

Our example highlights the importance of maintaining the efforts to link the different ontologies and databases, as systematically using universal identifiers to describe each physical entity rather than using strings, even if it means using an ID of a rather generic term. This will allow to better exploit the growing mass of data and studies [1]. In conclusion, this study opens new perspectives to integrate simultaneously multimodal data and find the systemic organization behind experimental data.

CHAPTER 7

DISCUSSION

We focused on analysis methods for complex phenotypes that are out of reach of traditional approaches. Our main hypothesis relied on the possibility to better represent relationships between molecules that regulate a phenotype by combining experimental results and public knowledge, and by adopting a holistic view on biological organization. We suggested that integrating different levels of omics as a whole will help understand biological systems, especially by considering the cascade of events and the interactions between entities. To this end, we focused on a complex phenotype: feed efficiency in growing pigs.

In this context of feed efficiency, we highlighted relationships between two types of biological entities: transcriptomics and metabolomics. We identified modules of co-expressed genes by constructing a weighted gene co-expression network and computing pairwise correlations among the molecular probes (mRNA expression levels) in the whole blood sampled from growing pigs belonging to divergent lines for feed efficiency. Next, we correlated the eigengene of each module with feed efficiency traits, leading to the identification of 8 modules of co-expressed genes associated with Feed Conversion Ration (FCR), the on-farm measure of feed efficiency. These modules were enriched in genes involved in immune and defense-related processes, responses to organic substance, ribosome biogenesis and translation, and cell development and learning processes. Furthermore, we summarized the metabolomics data, comprising levels of fatty acids and targeted metabolites in the whole blood of the same animals, and then correlated them with the eigengene of the modules to link metabolic profiles to the co-expressed genes related to feed efficiency. This allowed the identification of a module of co-expressed genes related to immune process associated with circulating concentrations of omega-3 fatty acids in the plasma.

Then, we addressed both the challenge of integrating experimental data and knowledge bases to bridge the gap between the molecular and cellular levels, as well as the challenge of analyzing data and extracting knowledge using graph-based metrics. To investigate co-regulation and to identify potential regulators for co-expressed gene modules previously associated with feed efficiency, we developed a knowledge-guided approach. This approach leverages

the BioPAX format as a robust framework for integrating transcriptomic and metabolomic data.

To reason at the BioPAX level, we had to address issues related to non-conformity and redundancy in biological databases stored in BioPAX format. We developed a method for detecting and correcting data inconsistencies related to molecular complexes, and we applied this methods to the widely-used Reactome pathway database. We successfully detected and corrected 39% of molecular complexes in the *Homo sapiens* version of the database. Similar levels of improvement, ranging from 30% (in *Plasmodium falciparum*) to 40% (in *Sus scrofa*, *Bos taurus*, *Canis familiaris*, and *Gallus gallus*), were observed for other species suggesting robust improvement. As an additional consequence, these corrections allowed to identify complex redundancies.

We focused mainly on the eight modules of co-expressed genes, specifically selected due to their strong associations with animal phenotypes. Among these eight modules, our method successfully retrieved a significant number of protein nodes from Reactome, ranging from 56% to 100% of the proteins corresponding to the genes in these modules. Furthermore, we directed our attention to the 50 metabolites quantified in plasma samples. We retrieved 49 ChEBI IDs out of the initial set of 50 targeted metabolite names, corresponding to 165 ChEBI elements when considering sub-elements and enantiomers. However, we found only 27 corresponding nodes within Reactome.

While RDF was suited to data integration and symbolic reasoning base on ontologies, we switched to Neo4j database to be able to perform more complex analysis based on the graph topology. We conducted a comparison between the shortest paths connecting participants in co-expression modules and those in randomly-generated modules of the same size. Among the modules studied, we observed two distinct behaviors: the biggest modules exhibited shorter paths and fewer biochemical reactions in the co-expression module compared to randomizations, while smaller ones did not show significant differences from random modules. Furthermore, we observe that co-expression modules are significantly more connected by small molecules than random modules. This finding validates the decision to use a metabolic network.

7.1 Benefits of our approaches

Dealing with BioPAX We have demonstrated that Semantic Web technologies can address the challenges of standardizing and improving data quality. Our approach, which relies on semantically-rich queries for identifying and fixing invalid complexes in BioPAX, is not lim-

ited to a single database but is reproducible on others. This enhances data conformity and the graph analysis by repairing its topology. Although this issue of non-compliance was noted by Stromback [95] in 2005, it was neither detected nor rectified by any existing methods, such as the BioPAX-validator designed to validate BioPAX files [105], nor by parsers such as BioPAX-Parser [157], Paxtools [110], or PAX2GraphML [158]. As a result, this allows to apply reasoning methods on higher-quality data, leading to a better understanding of the regulation of complex phenotypes.

In the scope of our project, we examined various approaches to leverage pathways datasets, particularly those within the Reactome database. One option could have been to work directly within Reactome using the providing Neo4j database or to export it to a Python data model, such as STARGate-X does [159]. However, this option would have been highly dependant on data and on Reactome's specific data model, as emphasized by the authors of STARGate-X: "to use our tool with a different pathway repository, it would be necessary to store Pathway data in the graph database according to Reactome's schema for labels and nodes.". With the aim of proposing generic methods and analyzes, we opted for the use of BioPAX, a standard format for representing and sharing pathways data. This choice allows our work to be data-independent and reproducible across various other pathways databases.

Due to its semantic richness, BioPAX is a complex format. One approach would have been to simplify its manipulation by transforming the data into a Python model. For instance, tools like PAX2GraphML, which converts BioPAX files into regulated reaction graphs in order to facilitates their handling [158]. However, I made the decision to stay as close to the original data as possible by delving into the complexity of the BioPAX ontology, thereby minimizing the risk of loss or excessive modification of information. Consequently, our approach preserves data integrity and allows us to propose generic methods. As a side-result, our work constitutes a concrete example of the utilization of BioPAX with SPARQL queries.

Heterogeneous data integration We demonstrated that SPARQL and Semantic Web technologies are highly effective and well-suited for integrating -omics data. The representation of this data in a graph format facilitates the application of multi-layered graph-based integration. This approach enables the combined use of various existing databases, knowledge resources, and ontologies through both local and public SPARQL endpoints. Furthermore, the BioPAX format proves to be suitable for integrating data across multiple levels of biological organization. As noted by Cavill et al. [27] and Eicher et al. [26], the integration of metabolomics data

with other -omics data poses a significant challenge due to the absence of a direct link between metabolites and genes, contrary to those existing between genes and proteins. Our integration strategy focuses on mapping biological entities of interest within a metabolic network. While the use of metabolic networks for data integration is not a novel concept [160], our approach distinguishes itself through its innovative nature. In contrast to the current state of the art, our approach involves the creation of a sort of multi-layered network that integrates knowledge across various levels, as illustrated in the schema integration Figure 7.1. We leverage the Semantic Web to enhance interoperability, reproducibility, and flexibility when applying our approach to address new scientific inquiries or adapting to changes in databases. Our primary objective is the development of reusable tools that can be applied to diverse datasets and scientific questions.

In addition, we integrated Semantic Web technologies with Neo4j using the Neosemantics plugin. This combination of technologies provides a robust framework for analyzing complex data, and we used it to perform graph traversal and exploration. Furthermore, we found that certain tasks are more efficiently accomplished using SPARQL, particularly integration and some aspects of reasoning. Conversely, Cypher is indispensable for tasks involving intricate graph traversal. Additionally, there are intermediate tasks that can be handled equally well using both SPARQL and Cypher.

We initially identified co-expressed gene modules using a traditional approach. Subsequently, using our integration method, we demonstrated that these genes tend to be closer and more interconnected in the metabolic graph than would be expected from a random distribution. This observation suggests potential co-regulation within these modules.

Our approach differentiates from dimensionality reduction integration methods by introducing an integrative approach that capitalizes on the biological knowledge context of the studied physical entities. Additionally, our approach stands distinct from strictly defined multiplex graph approaches (with common nodes across layers). These methods often face challenges in effectively representing biochemical reactions. Unlike these methods, our approach utilizes oriented and labeled graphs, providing a more nuanced and accurate representation of complex biological interactions, moving beyond the constraints of adjacency matrices.

Understanding feed efficiency Gilbert et al. suggested, based on a divergent selection experiment on Residual Feed Intake (RFI), a measure of feed efficiency, that pigs have various ways to achieve efficient use of feed [92]. Therefore, examining the biological bases of inter-individual differences in feed efficiency is of utmost interest.

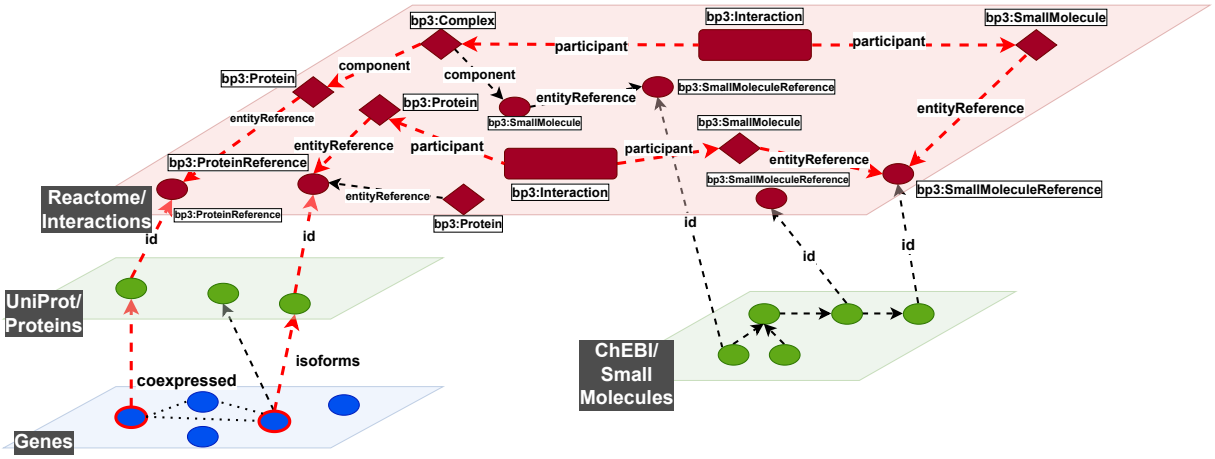

Figure 7.1 – Integration schema. Each layer represents a type of data from a data source linked to the other layers by different properties.

Our work revealed the existence of a small number of regulatory networks significantly associated with Feed Conversion Ratio (FCR). Whereas FCR is considered a production performance trait, the majority of the co-expressed gene networks were participating to immune process. The contribution of immunity to feed efficiency has been previously suggested through differential analyses between divergent groups for RFI at the transcriptomic level [97] and at the proteomic level [161]. We also demonstrate that some of the genes previously identified as having significant variation dynamics between divergent groups of pigs for RFI are important elements in these regulatory networks influencing inter-individual variability in feed efficiency.

The other benefit of our study is to establish connections between the transcriptome and metabolome, revealing links between immunity and fatty acid composition. This association had not previously been demonstrated and is relevant for future nutritional recommendations to obtain good synergy between production and health. Our study represents an initial effort toward unraveling the black box of complex traits, which remains a significant challenge in current research [125]. Other studies have also combined transcriptomic and metabolomic data for production traits other than feed efficiency, such as body adiposity. However, these studies often limit their analyses to separate omics levels [162, 163].

7.2 Limitations of the study

This work is subject to several limitations, including the incompleteness of biological knowledge, the modeling assumptions made, algorithmic limitations, semantic integration

issues, and identification challenge.

Firstly, it's important to acknowledge that public databases, much like our understanding of living organisms, are inherently incomplete. Consequently, in our investigations, we seek connections between our various entities of interest among the known interactions cataloged in these databases. Consequently, relying on public databases implies a limitation to well-studied pathways and functions. It is crucial to remain mindful that both databases and biological analyses, such as those employed in the context of feed efficiency applications, are intrinsically characterized by their incompleteness due to the vastness and complexity of biological systems.

Secondly, it's worth noting that we have made a simplifying assumption that a gene corresponds to one protein (or more), although reality is often more intricate. The presence of an expressed transcript does not necessarily imply the translation into the corresponding protein, as many factors can influence protein production. Regulatory steps, quality controls, and various other factors are involved in protein synthesis. Furthermore, we do not account for post-translational modifications, nor consider the notion of flux. Considering flux may be important for some biological questions involving the dynamic aspects of biological processes. Only the stoichiometric coefficients of the interactions stored in Reactome are available in the graph, and for now we do not use this information.

One of the main challenges when integrating experimental data with public knowledge is accurately identifying biological entities, especially because different sources use varying naming conventions or identifiers for the same entities. To overcome this challenge we utilized ontologies that offer unified names and identifiers, namely UniProt and ChEBI. Identifying proteins is relatively straightforward; however, special attention needs to be given to isoforms since there are multiple ways to identify them within both UniProt and Reactome. The complexity intensifies when dealing with metabolites. Indeed, the ChEBI ontology is notably semantically-rich, making it challenging to ensure a consistent understanding of specific biological entities in different context. For instance, a search for "lactate" in ChEBI yields 79 entries, encompassing not only `lactate` but also `(R)-lactate` and `(S)-lactate`. A similar challenge was encountered by Cavill et al. regarding lactate in the KEGG knowledge base [27]. They identified multiple identifiers for different forms of lactate; however, not all these lactate forms were involved in the pathways. Their conclusion emphasizes the need for individual cases to be evaluated manually to manage these overlapping metabolites and

ambiguous assignments effectively. Consequently, determining the precise identifier used in each study and selecting the identifier for storage in databases becomes intricate. Experimental methods often cannot identify elements as finely as the most specific terms in ChEBI. As a result, within Reactome, although the concept of `SmallMoleculeReference` exists in BioPAX, we encounter small molecules described at various levels within the ChEBI ontology. Indeed, `SmallMoleculeReference` nodes are designed to group different forms of a small molecule under a relatively generic term. However, we observed, for example in version 84, that `SmallMoleculeReference` number 266 possesses the ChEBI identifier ChEBI:57540 (NAD Anion) and yet serves as the reference entity for several `SmallMolecules` annotated "NAD+". To handle this challenge, we manually identified the generic terms corresponding to metabolites of interest. Subsequently, through a federated query, we searched in ChEBI for its descendant terms as well as potential enantiomers. These terms were then searched within Reactome for comprehensive integration.

A further limitation lies in our data model, which is generic with regard to the ontologies used but not optimized for existing graph traversal algorithms. In Neo4j, the graph being analyzed is a labeled property graph where edges are derived from the predicates of the BioPAX ontology, and node types of its classes. This implies that traditional graph mining algorithms cannot be directly applied. They need to be tailored to our data schema to be effectively utilized. The complexity of adaptation varies based on the specific algorithms we plan to employ.

Recognizing these challenges becomes a stepping stone toward innovative solutions and breakthroughs.

7.3 Perspectives and potential future improvement and research directions

In the immediate future of this research project, our focus is on refining our graph traversal methods. Currently, efforts are underway to avoid traversing through small molecules acting as hubs in the graph (such as water, H+, ATP, NAD, etc.). By excluding these molecules, we can avoid considering the shortest paths passing through these ubiquitous entities [72]. This exclusion will enable a reevaluation of the shortest paths computed in Chapter 6. Consequently, this

refinement might reveal more distinct patterns between co-expression modules and randomly generated modules. Furthermore, we aim to compare the list of metabolites present on these paths with those in our list of interest. Moreover, we intend to identify participants in the interactions connecting our proteins, aiming to systematically detect biological entities that might be related with the studied phenotypes.

Another planned step involves conducting various topological analyses on these specific sub-networks of interest. So far, our focus has been on identifying the shortest paths between two entities within sub-networks of interest. This decision was made to assess whether the sub-networks of proteins of interest are more interconnected than what would be expected by random chance. However, to identity potential regulators of the phenotype, it is important to consider that the shortest paths in metabolic networks are not always the most efficient or biologically relevant. Frainay and Jourdan highlighted the challenge of selecting meaningful paths, and the fact there is no one method more appropriate than others [72]. One idea worth exploring is to traverse the graph using alternative algorithms, such as random walks, where proteins within a module are chosen as seed nodes. This approach can help identify nodes that are frequently traversed, providing valuable insights into essential related biological entities. Another avenue for investigation involves identifying graph traversal algorithms specifically tailored to our biological questions, especially those capable of navigating multilayer networks. This approach could emphasize transitions between different types of biological entities. Tools like MuxViz offer various topological measures for multilayer graphs [70], prompting us to consider representing our graph as matrices to apply these measures effectively. In any cases, whether adapting the algorithm implementation or refining our graph representation, extensive work is required to ensure compatibility with our data model.

Enhancing entity identification is of primordial importance to exploit the large amount of data generated and stored in databases. Interpreting experimental data in the context of existing biological knowledge requires accurate referencing. Therefore, addressing these challenges is crucial. This can be achieved by systematically adopting existing standardized naming conventions, developing more efficient entity resolution algorithms, or exploring the use of machine learning techniques for entity disambiguation (as in [164]).

We introduced reproducible SPARQL queries for identifying and fixing invalid complexes in BioPAX databases, however as mentioned in the limitations, there are other types of inconsistencies that could potentially impact analyses. Further work is needed to identify these issues and to address them, it would thereby enhancing the overall quality and reliability of the

graph data.

A long-term perspective would be to enrich the existing graph with additional layers to capture diverse interactions or incorporate extra knowledge by linking other ontologies to the physical entities within the graph. In our specific study of feed efficiency, it could be valuable to integrate not only different -omics levels but also different tissues [165], such as muscles or less invasive ones like feces or milk in dairy species [134, 166]. This expansion would create a more interconnected representation and may enlighten new interesting paths. Given that we explore the Reactome database, it would also be valuable to reason at the level of pathways, enabling us to identify the specific ones in which the entities of interest are involved.

Feed efficiency is one example of complex phenotype were the simultaneous integration of different -omics datasets might be useful. There are many other complex phenotypes that could be investigated with our approach. Applying our approach to another experimental dataset or a different biological question will allow to validate and reinforce its applicability.

7.4 Conclusion

The methods devised in this thesis have wide-ranging applicability to address diverse biological questions related to complex phenotypes. Significant efforts have been dedicated to develop generic and reusable tools, both in terms of the selected techniques and their implementations. Furthermore, all codes have been made accessible through notebooks, ensuring their availability for future research endeavors.

Biological data exhibit a large and multifaceted heterogeneity that can be leveraged to gain intricate insights into biological systems. However, achieving this requires a collective effort to bridge the need for the development of integration tools and the use of a common language for denoting biological entities, enabling the seamless integration and interpretation of diverse data types.

BIBLIOGRAPHY

1. Stephens, Z. D. *et al.,* Big Data: Astronomical or Genomical?, en, *PLOS Biology* **13,** e1002195, ISSN: 1545-7885, https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002195 (2022) (July 2015).
2. Rigden, D. J. & Fernández, X. M., The 2021 Nucleic Acids Research database issue and the online molecular biology database collection, *Nucleic Acids Research* **49,** D1–D9, ISSN: 0305-1048, https://doi.org/10.1093/nar/gkaa1216 (2022) (Jan. 2021).
3. Schuster, S. C., Next-generation sequencing transforms today's biology, en, *Nature Methods* **5,** 16–18, ISSN: 1548-7105, https://www.nature.com/articles/nmeth1156 (2023) (Jan. 2008).
4. Manzoni, C. *et al.,* Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences, *Briefings in Bioinformatics* **19,** 286–302, ISSN: 1477-4054, https://doi.org/10.1093/bib/bbw114 (2023) (Mar. 2018).
5. Rao, M. S. *et al.,* Comparison of RNA-Seq and Microarray Gene Expression Platforms for the Toxicogenomic Evaluation of Liver From Short-Term Rat Toxicity Studies, *Frontiers in Genetics* **9,** ISSN: 1664-8021, https://www.frontiersin.org/articles/10.3389/fgene.2018.00636 (2023) (2019).
6. Varadi, M. & Velankar, S., The impact of AlphaFold Protein Structure Database on the fields of life sciences, eng, *Proteomics* **23,** e2200128, ISSN: 1615-9861 (Sept. 2023).
7. Simpkin, A. J. *et al.,* Tertiary structure assessment at CASP15, eng, *Proteins,* ISSN: 1097-0134 (Sept. 2023).
8. Wodak, S. J., Vajda, S., Lensink, M. F., Kozakov, D. & Bates, P. A., Critical Assessment of Methods for Predicting the 3D Structure of Proteins and Protein Complexes, eng, *Annual Review of Biophysics* **52,** 183–206, ISSN: 1936-1238 (May 2023).
9. Lensink, M. F. *et al.,* Prediction of protein assemblies, the next frontier: The CASP14-CAPRI experiment, eng, *Proteins* **89,** 1800–1823, ISSN: 1097-0134 (Dec. 2021).

10. Zamboni, N., Saghatelian, A. & Patti, G. J., Defining the Metabolome: Size, Flux, and Regulation, *Molecular cell* **58,** 699–706, ISSN: 1097-2765, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4831058/ (2023) (May 2015).

11. Gaul, D. A. *et al.,* Highly-accurate metabolomic detection of early-stage ovarian cancer, eng, *Scientific Reports* **5,** 16351, ISSN: 2045-2322 (Nov. 2015).

12. Benson, D. A. *et al.,* GenBank, eng, *Nucleic Acids Research* **41,** D36–42, ISSN: 1362-4962 (Jan. 2013).

13. The UniProt Consortium, UniProt: the Universal Protein Knowledgebase in 2023, *Nucleic Acids Research* **51,** D523–D531, ISSN: 0305-1048, https://doi.org/10.1093/nar/gkac1052 (2023) (Jan. 2023).

14. Clough, E. & Barrett, T., The Gene Expression Omnibus database, *Methods in molecular biology (Clifton, N.J.)* **1418,** 93–110, ISSN: 1064-3745, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4944384/ (2023) (2016).

15. wwPDB consortium, Protein Data Bank: the single global archive for 3D macromolecular structure data, *Nucleic Acids Research* **47,** D520–D528, ISSN: 0305-1048, https://doi.org/10.1093/nar/gky949 (2023) (Jan. 2019).

16. Gillespie, M. *et al.,* The reactome pathway knowledgebase 2022, *Nucleic acids research,* In press (2021).

17. Amberger, J., Bocchini, C. A., Scott, A. F. & Hamosh, A., McKusick's Online Mendelian Inheritance in Man (OMIM®), *Nucleic Acids Research* **37,** D793–D796, ISSN: 0305-1048, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2686440/ (2023) (Jan. 2009).

18. De Matos, P. *et al.,* ChEBI: a chemistry ontology and database, *Journal of cheminformatics* **2,** 1–1 (2010).

19. Sivade (Dumousseau), M. *et al.,* Encompassing new use cases - level 3.0 of the HUPO-PSI format for molecular interactions, en, *BMC Bioinformatics* **19,** 134, ISSN: 1471-2105, https://doi.org/10.1186/s12859-018-2118-1 (2023) (Apr. 2018).

20. The Gene Ontology Consortium *et al.,* The Gene Ontology knowledgebase in 2023, *Genetics* **224,** iyad031, ISSN: 1943-2631, https://doi.org/10.1093/genetics/iyad031 (2023) (May 2023).

21. Köhler, S. *et al.,* The Human Phenotype Ontology in 2017, *Nucleic Acids Research* **45,** D865–D876, ISSN: 0305-1048, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5210535/ (2023) (Jan. 2017).

22. *Livestock Ontologies – ATOL – EOL – AHOL* fr-FR, 2014, https://www.atol-ontology.com/ (2023).

23. Sargsyan, A. *et al.,* The COVID-19 Ontology, eng, *Bioinformatics (Oxford, England)* **36,** 5703–5705, ISSN: 1367-4811 (Apr. 2021).

24. Perez-Riverol, Y. *et al.,* Discovering and linking public omics data sets using the Omics Discovery Index, en, *Nature Biotechnology* **35,** 406–409, ISSN: 1546-1696, https://www.nature.com/articles/nbt.3790 (2023) (May 2017).

25. Haas, R. *et al.,* Designing and interpreting 'multi-omic' experiments that may change our understanding of biology, *Current Opinion in Systems Biology* **6,** 37–45, ISSN: 2452-3100, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7477987/ (2023) (Dec. 2017).

26. Eicher, T. *et al.,* Metabolomics and Multi-Omics Integration: A Survey of Computational Methods and Resources, *Metabolites* **10,** 202, ISSN: 2218-1989, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7281435/ (2023) (May 2020).

27. Cavill, R., Jennen, D., Kleinjans, J. & Briedé, J. J., Transcriptomic and metabolomic data integration, *Briefings in Bioinformatics* **17,** 891–901, ISSN: 1467-5463, https://doi.org/10.1093/bib/bbv090 (2023) (Sept. 2016).

28. Zhang, B. *et al.,* Proteogenomic characterization of human colon and rectal cancer, en, *Nature* **513,** 382–387, ISSN: 1476-4687, https://www.nature.com/articles/nature13438 (2023) (Sept. 2014).

29. Macaulay, I. C., Ponting, C. P. & Voet, T., Single-Cell Multiomics: Multiple Measurements from Single Cells, *Trends in Genetics* **33,** 155–168, ISSN: 0168-9525, https://www.sciencedirect.com/science/article/pii/S016895251630169X (2023) (Feb. 2017).

30. Kristensen, V. N. *et al.,* Principles and methods of integrative genomic analyses in cancer, en, *Nature Reviews Cancer* **14,** 299–313, ISSN: 1474-1768, https://www.nature.com/articles/nrc3721 (2023) (May 2014).

31. Menyhárt, O. & Gyorffy, B., Multi-omics approaches in cancer research with applications in tumor subtyping, prognosis, and diagnosis, *Computational and Structural Biotechnology Journal* **19,** 949–960, ISSN: 2001-0370, https://www.sciencedirect.com/science/article/pii/S2001037021000131 (2023) (Jan. 2021).

32. Weinstein, J. N. *et al.,* The Cancer Genome Atlas Pan-Cancer analysis project, en, *Nature Genetics* **45,** 1113–1120, ISSN: 1546-1718, https://www.nature.com/articles/ng.2764. (2023) (Oct. 2013).

33. Angione, C., Conway, M. & Lió, P., Multiplex methods provide effective integration of multi-omic data in genome-scale models, *BMC Bioinformatics* **17,** 83, ISSN: 1471-2105, https://doi.org/10.1186/s12859-016-0912-1 (2023) (Mar. 2016).

34. Rattray, N. J. W. *et al.,* Beyond genomics: understanding exposotypes through metabolomics, *Human Genomics* **12,** 4, ISSN: 1479-7364, https://doi.org/10.1186/s40246-018-0134-x (2023) (Jan. 2018).

35. Proctor, L. M. *et al.,* The Integrative Human Microbiome Project, en, *Nature* **569,** 641–648, ISSN: 1476-4687, https://www.nature.com/articles/s41586-019-1238-8 (2023) (May 2019).

36. Lloyd-Price, J. *et al.,* Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases, en, *Nature* **569,** 655–662, ISSN: 1476-4687, https://www.nature.com/articles/s41586-019-1237-9 (2023) (May 2019).

37. Cantini, L. *et al.,* Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer, eng, *Nature Communications* **12,** 124, ISSN: 2041-1723 (Jan. 2021).

38. Zhou, G., Li, S. & Xia, J., en, *in* (ed Li, S.) 469–487 (Springer US, New York, NY, 2020), ISBN: 9781071602393, https://doi.org/10.1007/978-1-0716-0239-3_23 (2023).

39. Subramanian, I., Verma, S., Kumar, S., Jere, A. & Anamika, K., Multi-omics Data Integration, Interpretation, and Its Application, en, *Bioinformatics and Biology Insights* **14,** 1177932219899051, ISSN: 1177-9322, https://doi.org/10.1177/1177932219899051 (2023) (Jan. 2020).

40. Yan, J., Risacher, S. L., Shen, L. & Saykin, A. J., Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data, *Briefings in Bioinformatics* **19,** 1370–1381, ISSN: 1467-5463, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6454489/ (2023) (June 2017).

41. Bersanelli, M. *et al.,* Methods for the integration of multi-omics data: mathematical aspects, *BMC Bioinformatics* **17,** 15, ISSN: 1471-2105, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4959355/ (2023) (Jan. 2016).

42. Shen, H. & Huang, J. Z., Sparse principal component analysis via regularized low rank matrix approximation, *Journal of Multivariate Analysis* **99,** 1015–1034, ISSN: 0047-259X, https://www.sciencedirect.com/science/article/pii/S0047259X07000887 (2023) (July 2008).

43. Rau, A. *et al.,* Individualized multi-omic pathway deviation scores using multiple factor analysis, *Biostatistics* **23,** 362–379, ISSN: 1465-4644, https://doi.org/10.1093/biostatistics/kxaa029 (2023) (Apr. 2022).

44. Barker, M. & Rayens, W., Partial least squares for discrimination, en, *Journal of Chemometrics* **17,** 166–173, ISSN: 1099-128X, https://onlinelibrary.wiley.com/doi/abs/10.1002/cem.785 (2023) (2003).

45. Ruiz-Perez, D., Guan, H., Madhivanan, P., Mathee, K. & Narasimhan, G., So you think you can PLS-DA?, *BMC Bioinformatics* **21,** 2, ISSN: 1471-2105, https://doi.org/10.1186/s12859-019-3310-7 (2023) (Dec. 2020).

46. Li, W., Zhang, S., Liu, C.-C. & Zhou, X. J., Identifying multi-layer gene regulatory modules from multi-dimensional genomic data, *Bioinformatics* **28,** 2458–2466, ISSN: 1367-4803, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3463121/ (2023) (Oct. 2012).

47. Singh, A. *et al.,* DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays, eng, *Bioinformatics (Oxford, England)* **35,** 3055–3062, ISSN: 1367-4811 (Sept. 2019).

48. Rohart, F., Eslami, A., Matigian, N., Bougeard, S. & Lê Cao, K.-A., MINT: a multivariate integrative method to identify reproducible molecular signatures across independent experiments and platforms, *BMC Bioinformatics* **18,** 128, ISSN: 1471-2105, https://doi.org/10.1186/s12859-017-1553-8 (2023) (Feb. 2017).

49. Poirier, S., Déjean, S., Midoux, C., Cao, K.-A. & Chapleur, O., Integrating independent microbial studies to build predictive models of anaerobic digestion inhibition by ammonia and phenol, *Bioresource Technology* **316,** 123952 (Aug. 2020).

50. Lee, D. & Seung, H. S., *Algorithms for Non-negative Matrix Factorization in Advances in Neural Information Processing Systems* (eds Leen, T., Dietterich, T. & Tresp, V.) **13** (MIT Press, 2000), https://proceedings.neurips.cc/paper_files/paper/2000/file/f9d1152547c0bde01830b7e8bd60024c-Paper.pdf.

51. Argelaguet, R. *et al.,* Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets, *Molecular systems biology* **14,** e8124 (2018).

52. Argelaguet, R. *et al.,* MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data, *Genome Biology* **21,** 111 (2020).

53. Velten, B. *et al.,* Identifying temporal and spatial patterns of variation from multimodal data using MEFISTO, en, *Nature Methods* **19,** 179–186, ISSN: 1548-7105, https://www.nature.com/articles/s41592-021-01343-9 (2023) (Feb. 2022).

54. Kmetzsch, V. *et al.,* Disease Progression Score Estimation From Multimodal Imaging and MicroRNA Data Using Supervised Variational Autoencoders, *IEEE Journal of Biomedical and Health Informatics* **26,** 6024–6035 (2022).

55. Reiman, D., Layden, B. T. & Dai, Y., MiMeNet: Exploring microbiome-metabolome relationships using neural networks, en, *PLOS Computational Biology* **17,** e1009021, ISSN: 1553-7358, https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1009021 (2023) (May 2021).

56. Langfelder, P. & Horvath, S., WGCNA: an R package for weighted correlation network analysis, *BMC Bioinformatics* **9,** 559, ISSN: 1471-2105, https://doi.org/10.1186/1471-2105-9-559 (2022) (Dec. 2008).

57. Reverter, A. & Chan, E. K. F., Combining partial correlation and an information theory approach to the reversed engineering of gene co-expression networks, *Bioinformatics* **24,** 2491–2497, ISSN: 1367-4803, eprint: https://academic.oup.com/bioinformatics/article-pdf/24/21/2491/49055106/bioinformatics_24_21_2491.pdf, https://doi.org/10.1093/bioinformatics/btn482 (Sept. 2008).

58. Levine, M. & Davidson, E. H., Gene regulatory networks for development, *Proceedings of the National Academy of Sciences* **102,** 4936–4942, https://www.pnas.org/doi/10.1073/pnas.0408031102 (2023) (Apr. 2005).

59. Brun, C. *et al.,* Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network, *Genome Biology* **5,** R6, ISSN: 1474-760X, https://doi.org/10.1186/gb-2003-5-1-r6 (2023) (Dec. 2003).

60. Becker, E., Robisson, B., Chapple, C. E., Guénoche, A. & Brun, C., Multifunctional proteins revealed by overlapping clustering in protein interaction network, *Bioinformatics* **28,** 84–90, ISSN: 1367-4803, eprint: https://academic.oup.com/bioinformatics/article-pdf/28/1/84/50568263/bioinformatics_28_1_84.pdf, https://doi.org/10.1093/bioinformatics/btr621 (Nov. 2011).

61. Bansal, P. *et al.,* Rhea, the reaction knowledgebase in 2022, *Nucleic Acids Research* **50,** D693–D700, ISSN: 0305-1048, https://doi.org/10.1093/nar/gkab1016 (2023) (Jan. 2022).

62. Kanehisa, M. & Goto, S., KEGG: kyoto encyclopedia of genes and genomes, *Nucleic acids research* **28,** 27–30 (2000).

63. Goh, K.-I. *et al.,* The human disease network, eng, *Proceedings of the National Academy of Sciences of the United States of America* **104,** 8685–8690, ISSN: 0027-8424 (May 2007).

64. Baker, M., Big biology: The 'omes puzzle, en, *Nature* **494,** 416–419, ISSN: 1476-4687, https://www.nature.com/articles/494416a (2023) (Feb. 2013).

65. Mitra, K., Carvunis, A.-R., Ramesh, S. K. & Ideker, T., Integrative approaches for finding modular structure in biological networks, *Nature reviews. Genetics* **14,** 719–732, ISSN: 1471-0056, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3940161/ (2023) (Oct. 2013).

66. Hartwell, L. H., Hopfield, J. J., Leibler, S. & Murray, A. W., From molecular to modular cell biology, en, *Nature* **402,** C47–C52, ISSN: 1476-4687, https://www.nature.com/articles/35011540 (2023) (Dec. 1999).

67. Yugi, K., Kubota, H., Hatano, A. & Kuroda, S., Trans-Omics: How To Reconstruct Biochemical Networks Across Multiple 'Omic' Layers, eng, *Trends in Biotechnology* **34,** 276–290, ISSN: 1879-3096 (Apr. 2016).

68. De Domenico, M. *et al.,* Mathematical Formulation of Multilayer Networks, *Phys. Rev. X* **3,** 041022, https://link.aps.org/doi/10.1103/PhysRevX.3.041022 (4 Dec. 2013).

69. Kivelä, M. *et al.,* Multilayer networks, *Journal of Complex Networks* **2,** 203–271, ISSN: 2051-1310, https://doi.org/10.1093/comnet/cnu016 (2023) (Sept. 2014).

70. De Domenico, M., Porter, M. A. & Arenas, A., MuxViz: a tool for multilayer analysis and visualization of networks, *Journal of Complex Networks* **3,** 159–176, ISSN: 2051-1310, eprint: https://academic.oup.com/comnet/article-pdf/3/2/159/1070864/cnu038.pdf, https://doi.org/10.1093/comnet/cnu038 (Oct. 2014).

71. Brohée, S. & van Helden, J., Evaluation of clustering algorithms for protein-protein interaction networks, *BMC Bioinformatics* **7,** 488, ISSN: 1471-2105, https://doi.org/10.1186/1471-2105-7-488 (2023) (Nov. 2006).

72. Frainay, C. & Jourdan, F., Computational methods to identify metabolic sub-networks based on metabolomic profiles, eng, *Briefings in Bioinformatics* **18,** 43–56, ISSN: 1477-4054 (Jan. 2017).

73. Nguyen, H. *et al.,* A Comprehensive Survey of Tools and Software for Active Subnetwork Identification, *Frontiers in Genetics* **10,** ISSN: 1664-8021, https://www.frontiersin.org/articles/10.3389/fgene.2019.00155 (2023) (2019).

74. Lazareva, O., Baumbach, J., List, M. & Blumenthal, D. B., On the limits of active module identification, *Briefings in Bioinformatics* **22,** bbab066, ISSN: 1477-4054, https://doi.org/10.1093/bib/bbab066 (2023) (Sept. 2021).

75. Köhler, S., Bauer, S., Horn, D. & Robinson, P. N., Walking the interactome for prioritization of candidate disease genes, eng, *American Journal of Human Genetics* **82,** 949–958, ISSN: 1537-6605 (Apr. 2008).

76. Carlin, D. E., Demchak, B., Pratt, D., Sage, E. & Ideker, T., Network propagation in the cytoscape cyberinfrastructure, eng, *PLoS computational biology* **13,** e1005598, ISSN: 1553-7358 (Oct. 2017).

77. Didier, G., Brun, C. & Baudot, A., Identifying communities from multiplex biological networks, *PeerJ* **3,** e1525, ISSN: 2167-8359, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4690346/ (2023) (Dec. 2015).

78. Bennett, L., Kittas, A., Muirhead, G., Papageorgiou, L. G. & Tsoka, S., Detection of Composite Communities in Multiplex Biological Networks, en, *Scientific Reports* **5,** 10345, ISSN: 2045-2322, https://www.nature.com/articles/srep10345 (2023) (May 2015).

79. Battiston, F., Nicosia, V. & Latora, V., Structural measures for multiplex networks, *Physical Review E* **89,** 032804, https://link.aps.org/doi/10.1103/PhysRevE.89.032804 (2023) (Mar. 2014).

80. Valdeolivas, A. *et al.,* Random walk with restart on multiplex and heterogeneous biological networks, eng, *Bioinformatics (Oxford, England)* **35,** 497–505, ISSN: 1367-4811 (Feb. 2019).

81. Baptista, A., Gonzalez, A. & Baudot, A., Universal multilayer network exploration by random walk with restart, en, *Communications Physics* **5,** 1–9, ISSN: 2399-3650, https://www.nature.com/articles/s42005-022-00937-9 (2023) (July 2022).

82. Novoa-del-Toro, E. M. *et al.,* A multi-objective genetic algorithm to find active modules in multiplex biological networks, en, *PLOS Computational Biology* **17,** e1009263, ISSN: 1553-7358, https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1009263 (2023) (Aug. 2021).

83. Berners Lee, T. *et al.,* A Framework for Web Science, *Foundations and Trends in Web Science* **1,** 1–130 (2007).

84. Shadbolt, N., Hall, W. & Berners Lee, T., The Semantic Web Revisited, *IEEE Intelligent Systems,* 96–101 (2006).

85. Stephens, S., LaVigna, D., DiLascio, M. & Luciano, J., Aggregation of Bioinformatics Data Using Semantic Web Technology, *Journal of Web Semantics* **4** (2006).

86. Sahoo, S. S., Bodenreider, O., Zeng, K. & Sheth, A., *An experiment in integrating large biomedical knowledge resources with RDF: Application to associating genotype and phenotype information in Proceedings of the WWW2007 Workshop on Health Care and Life Sciences Data Integration for the Semantic Web* (2007).

87. Gomez-Cabrero, D. *et al.,* Data integration in the era of omics: current and future challenges, *BMC systems biology* **8 Suppl 2,** I1 (2014).

88. Bizer, C., Heath, T. & Berners Lee, T., Linked Data–The story so far, *International Journal on Semantic Web and Information Systems* **5,** 1–22 (2009).

89. Larriba-Pey, J. L., Martínez-Bazán, N. & Domínguez-Sal, D., en, *in* (eds Koubarakis, M. *et al.*) 171–194 (Springer International Publishing, Cham, 2014), ISBN: 9783319105871, https://doi.org/10.1007/978-3-319-10587-1_4 (2023).

90. Francis, N. *et al., Cypher: An Evolving Query Language for Property Graphs in* (May 2018), 1433–1445.

91. Koch, R. M., Swiger, L. A., Chambers, D. & Gregory, K. E., Efficiency of Feed Use in Beef Cattle, *Journal of Animal Science* **22,** 486–494, ISSN: 0021-8812, https://doi.org/10.2527/jas1963.222486x (2023) (May 1963).

92. Gilbert, H. *et al.,* Review: divergent selection for residual feed intake in the growing pig, en, *animal* **11,** 1427–1439, ISSN: 1751-7311, 1751-732X, https://www.cambridge.org/core/journals/animal/article/review-divergent-selection-for-residual-feed-intake-in-the-growing-pig/E6D8E10772102377D22B8010F56FD906 (2022) (Sept. 2017).

93. Messad, F., Louveau, I., Renaudeau, D., Gilbert, H. & Gondret, F., Analysis of merged whole blood transcriptomic datasets to identify circulating molecular biomarkers of feed efficiency in growing pigs, *BMC Genomics* **22,** 501, ISSN: 1471-2164, https://doi.org/10.1186/s12864-021-07843-4 (2022) (July 2021).

94. Carmelo, V. A. O., Banerjee, P., da Silva Diniz, W. J. & Kadarmideen, H. N., Metabolomic networks and pathways associated with feed efficiency and related-traits in Duroc and Landrace pigs, en, *Scientific Reports* **10,** 255, ISSN: 2045-2322, https://www.nature.com/articles/s41598-019-57182-4 (2022) (Jan. 2020).

95. Strömbäck, L. & Lambrix, P., Representations of molecular pathways: an evaluation of SBML, PSI MI and BioPAX, eng, *Bioinformatics (Oxford, England)* **21,** 4401–4407, ISSN: 1367-4803 (Dec. 2005).

96. Demir, E. *et al.,* The BioPAX community standard for pathway data sharing, *Nature biotechnology* **28,** 935–942 (2010).

97. Jégou, M. *et al.,* Whole Blood Transcriptomics Is Relevant to Identify Molecular Changes in Response to Genetic Selection for Feed Efficiency and Nutritional Status in the Pig, en, *PLOS ONE* **11,** e0146550, ISSN: 1932-6203, https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0146550 (2022) (Jan. 2016).

98. Gondret, F. *et al.,* Dietary energy sources affect the partition of body lipids and the hierarchy of energy metabolic pathways in growing pigs differing in feed efficiency1,2, *Journal of Animal Science* **92,** 4865–4877, ISSN: 0021-8812, https://doi.org/10.2527/jas.2014-7995 (2023) (Nov. 2014).

99. Hucka, M. *et al.,* Evolving a lingua franca and associated software infrastructure for computational systems biology: the Systems Biology Markup Language (SBML) project, en, *Syst. Biol. (Stevenage)* **1,** 41–53 (June 2004).

100. Hucka, M. *et al.,* The Systems Biology Markup Language (SBML): Language specification for Level 3 Version 2 Core Release 2, en, *J. Integr. Bioinform.* **16** (June 2019).

101. Fearnley, L. G., Davis, M. J., Ragan, M. A. & Nielsen, L. K., Extracting reaction networks from databases-opening Pandora's box, eng, *Briefings in Bioinformatics* **15,** 973–983, ISSN: 1477-4054 (Nov. 2014).

102. Zahiri, J. *et al.,* Protein complex prediction: A survey, *Genomics* **112,** 174–183 (2019).

103. Spirin, V. & Mirny, L. A., Protein complexes and functional modules in molecular networks, *Proceedings of the National Academy of Sciences of the United States of America* **100,** 12123–12128 (2003).

104. Kazemzadeh, L., Deus, H., Dumontier, M. & Barry, F., Looking into Reactome through BioPAX lens, English, *CEUR Workshop Proceedings* **1054,** 37–40, ISSN: 1613-0073 (2013).

105. Rodchenkov, I., Demir, E., Sander, C. & Bader, G. D., The BioPAX Validator, eng, *Bioinformatics (Oxford, England)* **29,** 2659–2660, ISSN: 1367-4811 (Oct. 2013).

106. Meldal, B. H. M. *et al.,* Complex Portal 2022: new curation frontiers, *Nucleic acids research* **50,** D578–D586 (2022).

107. Caspi, R. *et al.,* The MetaCyc database of metabolic pathways and enzymes-a 2019 update, *Nucleic acids research* **48,** D445–D453 (2020).

108. Cerami, E. G. *et al.,* Pathway Commons, a web resource for biological pathway data, *Nucleic acids research* **39,** D685–D690 (2010).

109. Martens, M. *et al.,* WikiPathways: connecting communities, *Nucleic Acids Research* **49,** D613–D621, ISSN: 0305-1048, eprint: https://academic.oup.com/nar/article-pdf/49/D1/D613/35364599/gkaa1024.pdf, https://doi.org/10.1093/nar/gkaa1024 (Nov. 2020).

110. Demir, E. *et al.,* Using biological pathway data with paxtools, *PLoS computational biology* **9,** e1003194 (2013).

111. Gyori, B. M. & Hoyt, C. T., PyBioPAX: biological pathway exchange in Python, *Journal of Open Source Software* **7,** 4136, https://doi.org/10.21105/joss.04136 (2022).

112. Wang, Z., He, Y. & Tan, Z., Transcription Analysis of Liver and Muscle Tissues from Landrace Finishing Pigs with Different Feed Conversion Ratios, en, *Genes* **13,** 2067, ISSN: 2073-4425, https://www.mdpi.com/2073-4425/13/11/2067 (2022) (Nov. 2022).

113. Cantalapiedra-Hijar, G. *et al.,* Review: Biological determinants of between-animal variation in feed efficiency of growing beef cattle, en, *animal* **12,** s321–s335, ISSN: 1751-7311, 1751-732X, https://www.cambridge.org/core/journals/animal/article/review-biological-determinants-of-betweenanimal-variation-in-feed-efficiency-of-growing-beef-cattle/E14DE2A245113A54D1AD774C801D2427 (2023) (Dec. 2018).

114. Taiwo, G. A., Idowu, M., Denvir, J., Cervantes, A. P. & Ogunade, I. M., Identification of Key Pathways Associated With Residual Feed Intake of Beef Cattle Based on Whole Blood Transcriptome Data Analyzed Using Gene Set Enrichment Analysis, *Frontiers in Veterinary Science* **9,** ISSN: 2297-1769, https://www.frontiersin.org/articles/10.3389/fvets.2022.848027 (2022) (2022).

115. Schmidt, M. *et al.,* The Human Blood Transcriptome in a Large Population Cohort and Its Relation to Aging and Health, eng, *Frontiers in Big Data* **3,** 548873, ISSN: 2624-909X (2020).

116. Liu, H., Nguyen, Y. T., Nettleton, D., Dekkers, J. C. M. & Tuggle, C. K., Post-weaning blood transcriptomic differences between Yorkshire pigs divergently selected for residual feed intake, *BMC Genomics* **17,** 73, ISSN: 1471-2164, https://doi.org/10.1186/s12864-016-2395-x (2023) (Jan. 2016).

117. Salleh, S. M., Mazzoni, G., Løvendahl, P. & Kadarmideen, H. N., Gene co-expression networks from RNA sequencing of dairy cattle identifies genes and pathways affecting feed efficiency, *BMC Bioinformatics* **19,** 513, ISSN: 1471-2105, https://doi.org/10.1186/s12859-018-2553-z (2022) (Dec. 2018).

118. Banerjee, P., Carmelo, V. A. O. & Kadarmideen, H. N., Integrative Analysis of Metabolomic and Transcriptomic Profiles Uncovers Biological Pathways of Feed Efficiency in Pigs, en, *Metabolites* **10,** 275, ISSN: 2218-1989, https://www.mdpi.com/2218-1989/10/7/275 (2023) (July 2020).

119. Jégou, M. *et al.,* NMR-based metabolomics highlights differences in plasma metabolites in pigs exhibiting diet-induced differences in adiposity, eng, *European Journal of Nutrition* **55,** 1189–1199, ISSN: 1436-6215 (Apr. 2016).

120. Ghazalpour, A. *et al.,* Integrating Genetic and Network Analysis to Characterize Genes Related to Mouse Weight, en, *PLoS Genetics* **2** (ed Gibson, G.) e130, ISSN: 1553-7404, https://dx.plos.org/10.1371/journal.pgen.0020130 (2023) (Aug. 2006).

121. Zhang, B. & Horvath, S., A general framework for weighted gene co-expression network analysis, eng, *Statistical Applications in Genetics and Molecular Biology* **4,** Article17, ISSN: 1544-6115 (2005).

122. Yip, A. M. & Horvath, S., Gene network interconnectedness and the generalized topological overlap measure, en, *BMC Bioinformatics* **8,** 22, ISSN: 1471-2105, https://doi.org/10.1186/1471-2105-8-22 (2021) (Jan. 2007).

123. Lê, S., Josse, J. & Husson, F., FactoMineR: an R package for multivariate analysis, *Journal of statistical software* **25,** 1–18 (2008).

124. Kassambara, A. & Mundt, F., Package 'factoextra', *Extract and visualize the results of multivariate data analyses* **76** (2017).

125. Christensen, O. F., Börner, V., Varona, L. & Legarra, A., Genetic evaluation including intermediate omics features, *Genetics* **219,** iyab130, ISSN: 1943-2631, eprint: https://academic.oup.com/genetics/article-pdf/219/2/iyab130/41280101/iyab130.pdf, https://doi.org/10.1093/genetics/iyab130 (Aug. 2021).

126. Taiwo, G. *et al.,* Chemical Group-Based Metabolome Analysis Identifies Candidate Plasma Biomarkers Associated With Residual Feed Intake in Beef Steers, *Frontiers in Animal Science* **2,** ISSN: 2673-6225, https://www.frontiersin.org/articles/10.3389/fanim.2021.783314 (2023) (2022).

127. Goldansaz, S. A. *et al.,* Candidate serum metabolite biomarkers of residual feed intake and carcass merit in sheep, *Journal of Animal Science* **98,** skaa298, ISSN: 0021-8812, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7751155/ (2023) (Sept. 2020).

128. Hudson, N. J., Dalrymple, B. P. & Reverter, A., Beyond differential expression: the quest for causal mutations and effector molecules, *BMC Genomics* **13,** 356, ISSN: 1471-2164, https://doi.org/10.1186/1471-2164-13-356 (2023) (July 2012).

129. Alexandre, P. A. *et al.,* Liver transcriptomic networks reveal main biological processes associated with feed efficiency in beef cattle, *BMC Genomics* **16,** 1073, ISSN: 1471-2164, https://doi.org/10.1186/s12864-015-2292-8 (2023) (Dec. 2015).

130. Cho, D.-Y., Kim, Y.-A. & Przytycka, T. M., Chapter 5: Network Biology Approach to Complex Diseases, en, *PLOS Computational Biology* **8,** e1002820, ISSN: 1553-7358, https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002820 (2023) (Dec. 2012).

131. Horodyska, J. *et al.,* Analysis of meat quality traits and gene expression profiling of pigs divergent in residual feed intake, eng, *Meat Science* **137,** 265–274, ISSN: 1873-4138 (Mar. 2018).

132. Gondret, F. *et al.,* A transcriptome multi-tissue analysis identifies biological pathways and genes associated with variations in feed efficiency of growing pigs, *BMC Genomics* **18,** 244, ISSN: 1471-2164, https://doi.org/10.1186/s12864-017-3639-0 (2023) (Mar. 2017).

133. Cao, S. *et al.,* Reduced Meal Frequency Decreases Fat Deposition and Improves Feed Efficiency of Growing-Finishing Pigs, eng, *Animals: an open access journal from MDPI* **12,** 2557, ISSN: 2076-2615 (Sept. 2022).

134. Suárez-Vega, A. *et al.,* Feed efficiency in dairy sheep: An insight from the milk transcriptome, eng, *Frontiers in Veterinary Science* **10,** 1122953, ISSN: 2297-1769 (2023).

135. Labussière, E. *et al.,* Effect of inflammation stimulation on energy and nutrient utilization in piglets selected for low and high residual feed intake, eng, *Animal: An International Journal of Animal Bioscience* **9,** 1653–1661, ISSN: 1751-732X (Oct. 2015).

136. Rodrigues, L. A., Ferreira, F. N. A., Costa, M. O., Wellington, M. O. & Columbus, D. A., Factors affecting performance response of pigs exposed to different challenge models: a multivariate approach, *Journal of Animal Science* **99,** skab035, ISSN: 1525-3163, https://doi.org/10.1093/jas/skab035 (2023) (June 2021).

137. Alshabi, A. M., Vastrad, B., Shaikh, I. A. & Vastrad, C., Identification of Crucial Candidate Genes and Pathways in Glioblastoma Multiform by Bioinformatics Analysis, *Biomolecules* **9,** 201, ISSN: 2218-273X, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6571969/ (2023) (May 2019).

138. Bunter, K. L., Cai, W., Johnston, D. J. & Dekkers, J. C. M., Selection to reduce residual feed intake in pigs produces a correlated response in juvenile insulin-like growth factor-I concentration1, *Journal of Animal Science* **88,** 1973–1981, ISSN: 0021-8812, https://doi.org/10.2527/jas.2009-2445 (2023) (June 2010).

139. Do, D. N., Strathe, A. B., Ostersen, T., Pant, S. D. & Kadarmideen, H. N., Genome-wide association and pathway analysis of feed efficiency in pigs reveal candidate genes and pathways for residual feed intake, eng, *Frontiers in Genetics* **5,** 307, ISSN: 1664-8021 (2014).

140. Hou, Y. *et al.,* Neuronal Signal Transduction-Involved Genes in Pig Hypothalamus Affect Feed Efficiency as Revealed by Transcriptome Analysis, eng, *BioMed Research International* **2018,** 5862571, ISSN: 2314-6141 (2018).

141. Widmann, P. *et al.,* Systems biology analysis merging phenotype, metabolomic and genomic data identifies Non-SMC Condensin I Complex, Subunit G (NCAPG) and cellular maintenance processes as major contributors to genetic variability in bovine feed efficiency, eng, *PloS One* **10,** e0124574, ISSN: 1932-6203 (2015).

142. Gutiérrez, S., Svahn, S. L. & Johansson, M. E., Effects of Omega-3 Fatty Acids on Immune Cells, *International Journal of Molecular Sciences* **20,** 5028, ISSN: 1422-0067, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6834330/ (2023) (Oct. 2019).

143. Christ, A., Lauterbach, M. & Latz, E., Western Diet and the Immune System: An Inflammatory Connection, eng, *Immunity* **51,** 794–811, ISSN: 1097-4180 (Nov. 2019).

144. Morell, P. & Fiszman, S., Revisiting the role of protein-induced satiation and satiety, en, *Food Hydrocolloids, 30th anniversary special issue* **68,** 199–210, ISSN: 0268-005X, https://www.sciencedirect.com/science/article/pii/S0268005X1630340X (2023) (July 2017).

145. Argelaguet, R. *et al.,* Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets, *Molecular systems biology* **14,** e8124 (2018).

146. Rigden, D. J. & Fernández, X. M., The 2022 Nucleic Acids Research database issue and the online molecular biology database collection, *Nucleic acids research* **50,** D1–D10 (2022).

147. Kamdar, M. R. & Musen, M. A., An empirical meta-analysis of the life sciences linked open data on the web, *Scientific data* **8,** 24 (2021).

148. The UniProt Consortium, UniProt: the universal protein knowledgebase in 2021, *Nucleic Acids Research* **49,** D480–D489, ISSN: 0305-1048, (2023) (Jan. 2021).

149. Hastings, J. *et al.,* ChEBI in 2016: Improved services and an expanding collection of metabolites, *Nucleic Acids Research* **44,** D1214–D1219, ISSN: 0305-1048 (Oct. 2015).

150. Bairoch, A. *et al.,* The Universal Protein Resource (UniProt), *Nucleic Acids Research* **33,** D154–D159 (Jan. 2005).

151. Consortium, G. O., The gene ontology resource: 20 years and still GOing strong, *Nucleic acids research* **47,** D330–D338 (2019).

152. Juigné, C. *et al.,* Fixing molecular complexes in BioPAX standards to enrich interactions and detect redundancies using semantic web technologies, *Bioinformatics* **39,** btad257, ISSN: 1367-4811, (2023) (May 2023).

153. Juigné, C. *et al., Small networks of expressed genes in the whole blood and relationships to profiles in circulating metabolites provide insights in inter-individual variability of feed efficiency in growing pigs (preprint)* working paper or preprint, Mar. 2023, https://hal.science/hal-04112110.

154. Lowe, J. W. E., Humanising and dehumanising pigs in genomic and transplantation research, en, *Hist. Philos. Life Sci.* **44,** 66 (Nov. 2022).

155. Van Milgen, J., Modeling biochemical aspects of energy metabolism in mammals, eng, *The Journal of Nutrition* **132,** 3195–3202, ISSN: 0022-3166 (Oct. 2002).

156. Del Toro, N. *et al.,* The IntAct database: efficient access to fine-grained molecular interaction data, *Nucleic Acids Research* **50,** D648–D653, ISSN: 0305-1048, https://doi.org/10.1093/nar/gkab1006 (2023) (Jan. 2022).

157. Agapito, G., Pastrello, C., Guzzi, P. H., Jurisica, I. & Cannataro, M., BioPAX-Parser: parsing and enrichment analysis of BioPAX pathways, *Bioinformatics* **36,** 4377–4378, ISSN: 1367-4803, https://doi.org/10.1093/bioinformatics/btaa529 (2023) (Aug. 2020).

158. Moreews, F., Simon, H., Siegel, A., Gondret, F. & Becker, E., PAX2GRAPHML: a python library for large-scale regulation network analysis using BioPAX, *Bioinformatics* **37,** 4889–4891, ISSN: 1367-4803, https://doi.org/10.1093/bioinformatics/btab441 (2023) (Dec. 2021).

159. Marino, A., Sinaimeri, B., Tronci, E. & Calamoneri, T., STARGATE-X: a Python package for statistical analysis on the REACTOME network, eng, *Journal of Integrative Bioinformatics,* ISSN: 1613-4516 (Sept. 2023).

160. Blavy, P., Gondret, F., Lagarrigue, S., van Milgen, J. & Siegel, A., Using a large-scale knowledge database on reactions and regulations to propose key upstream regulators of various sets of molecules participating in cell metabolism, *BMC Systems Biology* **8,** 32, ISSN: 1752-0509, https://doi.org/10.1186/1752-0509-8-32 (2023) (Mar. 2014).

161. Zhu, S. *et al.,* Comparative Serum Proteome Analysis Indicates a Negative Correlation between a Higher Immune Level and Feed Efficiency in Pigs, en, *Veterinary Sciences* **10,** 338, ISSN: 2306-7381, https://www.mdpi.com/2306-7381/10/5/338 (2023) (May 2023).

162. Yu, H. *et al.,* Integrative Analysis of Blood Transcriptomics and Metabolomics Reveals Molecular Regulation of Backfat Thickness in Qinchuan Cattle, en, *Animals* **13,** 1060, ISSN: 2076-2615, https://www.mdpi.com/2076-2615/13/6/1060 (2023) (Jan. 2023).

163. Valdés-Hernández, J. *et al.,* Global analysis of the association between pig muscle fatty acid composition and gene expression using RNA-Seq, en, *Scientific Reports* **13,** 535, ISSN: 2045-2322, https://www.nature.com/articles/s41598-022-27016-x (2023) (Jan. 2023).

164. Galeota, E., Kishore, K. & Pelizzola, M., Ontology-driven integrative analysis of omics data through Onassis, en, *Scientific Reports* **10,** 703, ISSN: 2045-2322, https://www.nature.com/articles/s41598-020-57716-1 (2023) (Jan. 2020).

165. Ribeiro, G. *et al.,* Detection of potential functional variants based on systems-biology: the case of feed efficiency in beef cattle, *BMC Genomics* **23,** 774, ISSN: 1471-2164, https://doi.org/10.1186/s12864-022-08958-y (2023) (Nov. 2022).

166. Wu, J. *et al.,* Using nontargeted LC-MS metabolomics to identify the Association of Biomarkers in pig feces with feed efficiency, *Porcine Health Management* **7,** 39, ISSN: 2055-5660, https://doi.org/10.1186/s40813-021-00219-w (2023) (June 2021).

ANNEXES



Towards a reproducible interactome: semantic-based detection of redundancies to unify protein-protein interaction databases

Marc Melkonian, Camille Juigné, Olivier Dameron, Gwenaël Rabut, Emmanuelle Becker

▶ **To cite this version:**

Marc Melkonian, Camille Juigné, Olivier Dameron, Gwenaël Rabut, Emmanuelle Becker. Towards a reproducible interactome: semantic-based detection of redundancies to unify protein-protein interaction databases. Bioinformatics, Oxford University Press (OUP), 2022, pp.1-7. 10.1093/bioinformatics/btac013 . hal-03522989

HAL Id: hal-03522989

https://hal.archives-ouvertes.fr/hal-03522989

Submitted on 12 Jan 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Towards a reproducible interactome: semantic-based detection of redundancies to unify protein-protein interaction databases

Marc Melkonian, Camille Juigné, Olivier Dameron, Gwenaël Rabut, Emmanuelle Becker

▶ **To cite this version:**

Marc Melkonian, Camille Juigné, Olivier Dameron, Gwenaël Rabut, Emmanuelle Becker. Towards a reproducible interactome: semantic-based detection of redundancies to unify protein-protein interaction databases. Bioinformatics, Oxford University Press (OUP), 2022, 10.1093/bioinformatics/btac013 . hal-03522989

HAL Id: hal-03522989

https://hal.archives-ouvertes.fr/hal-03522989

Submitted on 12 Jan 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Databases and Ontologies

Towards a reproducible interactome: semantic-based detection of redundancies to unify protein-protein interaction databases

Marc Melkonian [1,2], Camille Juigné [1,3], Olivier Dameron[1], Gwenaël Rabut [2,*] and Emmanuelle Becker [1,*]

[1] Univ Rennes, Inria, CNRS, IRISA, F-35000, Rennes, France
[2] Univ Rennes, CNRS, IGDR - UMR 6290, F-35000, Rennes, France
[3] Pegase, Inrae, Institut Agro, 35590 Saint-Gilles, France.

*To whom correspondence should be addressed, equal contribution.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Information on protein-protein interactions is collected in numerous primary databases with their own curation process. Several meta-databases aggregate primary databases to provide more exhaustive datasets. In addition to exhaustivity, aggregation contributes to reliability by providing an overview of the various studies and detection methods supporting an interaction. However, interactions listed in different primary databases are partly redundant because some publications reporting protein-protein interactions have been curated by multiple primary databases. Mere aggregation can thus introduce a bias if these redundancies are not identified and eliminated. To overcome this bias, meta-databases rely on the Molecular Interaction ontology that describes interaction detection methods, but they do not fully take advantage of the ontology's rich semantics, which leads to systematically overestimating interaction reproducibility.

Results: We propose a precise definition of explicit and implicit redundancy, and show that both can be easily detected using Semantic Web technologies. We apply this process to a dataset from the APID meta-database and show that while explicit redundancies were detected by the APID aggregation process, about 15% of APID entries are implicitly redundant and should not be taken into account when presenting confidence-related metrics. More than 90% of implicit redundancies result from the aggregation of distinct primary databases, while the remaining occurs between entries of a single database. Finally, we build a "reproducible interactome" with interactions that have been reproduced by multiple methods or publications. The size of the reproducible interactome is drastically impacted by removing redundancies for both yeast (-59%) and human (-56%), and we show that this is largely due to implicit redundancies.

Availability: Software, data and results are available at `https://gitlab.com/nnet56/reproducible-interactome`, `https://reproducible-interactome.genouest.org/`, Zenodo (doi:10.5281/zenodo.5595037) and NDEx (doi:10.18119/N94302, doi:10.18119/N97S4D

Contact: emmanuelle.becker@irisa.fr, gwenael.rabut@inserm.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Protein-protein interactions (PPIs) play an ubiquitous and fundamental role in all biological processes. Description of PPIs is essential to understand how proteins operate at the molecular level and the construction of accurate

and comprehensive protein interaction networks (or interactomes) is an important aim of biological research (Bonetta, 2010; Cafarelli *et al.*, 2017; Luck *et al.*, 2020; Huttlin *et al.*, 2021).

PPIs can be probed using numerous interaction detection methods (IDMs), following biophysical (e.g. x-ray crystallography), biochemical (e.g. affinity purification) or genetic approaches (e.g. yeast two-hybrid). Importantly, since different IDMs probe PPIs in a different manner, they produce complementary results that often do not fully overlap. For instance, some IDMs are designed to detect binary interactions of proteins probed in pairs (e.g. yeast two-hybrid), while others probe interactions of protein groups assembled in complexes (e.g. affinity purification). Consequently, the biological interpretation of PPI networks depends on the underlying IDMs that have been used to produce them. Moreover, since IDMs can generate false positive and false negative interactions, multiple observations of a given PPI with different experimental techniques reinforce the confidence in this PPI. Accurate IDM annotation and interpretation is thus an important issue in interactome studies.

Information on published PPIs is collected in primary databases such as IntAct (Kerrien *et al.*, 2012), MINT (Calderone *et al.*, 2020), BioGRID (Oughtred *et al.*, 2019), DIP (Salwinski *et al.*, 2004) or HPRD (Keshava Prasad *et al.*, 2009). The major databases report IDMs using a controlled vocabulary defined by the Proteomics Standards Initiative-Molecular Interactions (PSI-MI) consortium (Sivade Dumousseau *et al.*, 2018). This vocabulary is structured in an ontology to represent the hierarchical relationships between IDM families by a directed acyclic graph.

Each primary database follows its own curation process with different literature mining, filtering, and reporting techniques. To address the resulting need for integration, several meta-databases aggregate information from multiple primary databases to provide more exhaustive PPI datasets. Some of these meta-databases, such as the Agile Protein Interactomes DataServer (APID) (Alonso-López *et al.*, 2016; Alonso-López *et al.*, 2019), HINT (Das and Yu, 2012) or mentha (Calderone *et al.*, 2013), focus exclusively on experimentally determined PPIs, while others, such as IID (Kotlyar *et al.*, 2019) or STRING (Szklarczyk *et al.*, 2019) also integrate predicted interactions, text mining results or other information.

The accurate aggregation of PPIs from multiple and partly redundant sources is not a trivial task (Turinsky *et al.*, 2010; Klapa *et al.*, 2013). Although the primary databases refer to the PSI-MI ontology, they do not necessarily select identical terms to annotate PPIs (Alonso-López *et al.*, 2019). Hence, a PPI observed in a single experiment reported in a given publication can be annotated with distinct IDM terms in different primary databases. Such annotation differences are usually not taken into account or corrected during the aggregation process.

APID, which unifies data from five of the largest PPI databases (Alonso-López *et al.*, 2016; Alonso-López *et al.*, 2019), implements an integration method that takes redundancy into account and enables to distinguish *'experimental evidences'* (i.e experimental observations reported in publications) from *'curation events'* (i.e. entries in PPI databases). For a given protein pair, multiple entries annotated with identical IDM and identical PubMed publication identifier (PMID) are considered as duplicates and counted as a single experimental evidence. In addition, IDMs are classified into *'binary'* and *'indirect'* methods and IDMs corresponding to related binary methods (e.g. 'two hybrid array' and 'two hybrid pooling approach') are assigned a common method type (e.g. 'two hybrid'). This common method type is then used instead of the original IDM to identify duplicate entries across multiple databases. This custom integration process is not fully satisfying since it is restricted to binary interactions and it does not take advantage of the PSI-MI ontology.

We propose a novel approach to integrate PPI information from primary databases. We define the conventional **explicit redundancy** and extend it with **implicit redundancy** based on parent-related terms in the PSI-MI ontology. We present a method relying on Semantic Web technologies that successfully detects and reconciles implicit redundancies in curation events compiled from multiple primary databases, opening the way to an improved automated curation process. Once curated for both explicit and implicit redundancies, the integrated set of experimental evidences can be used to determine the reproducible interactome supported by multiple experiments.

2 Approach

2.1 Explicit and implicit redundancy

Let us consider a pair of proteins (A, B) and count the number of non-redundant experiments reporting their interaction.

Primary databases such as BioGRID or IntAct can provide several entries corresponding to this protein pair. Usually, these entries differ in the IDM, the PMID, or both. An entry in these databases can thus be defined by a quadruplet

$$(A, B, M_i, P_x)$$

where A and B are the proteins, M_i is the IDM (such as 'affinity chromatography technology', 'anti-tag coimmunoprecipitation' or 'two hybrid', for the most frequent ones), and P_x is the PMID of the original article describing their interaction. When two entries only differ in the IDM, this should signify that the original article has observed the interaction using several experimental techniques. When two entries only differ in the PMID, this should signify that the interaction has been reproduced in two distinct studies using the same detection method.

For meta-databases such as APID, populated by aggregating curation events from other databases, an entry can be defined by a quintuplet

$$(A, B, M_i, P_x, D_a)$$

where D_a indicates the primary database indexing the interaction. Meta-databases can contain different types of redundancies:

- **Explicit redundancy** occurs when distinct entries referring to the same protein pair (A, B) and the same PMID P_x have an identical IDM M_i. This happens when two primary databases registered the same experimental evidence using the same IDM term. Explicit redundancies are detected and unified by APID and other meta-databases.
- **Implicit redundancy** occurs when distinct entries referring to the same protein pair and the same PMID have been annotated with different IDMs although they correspond to the same experimental evidence. In practice, this occurs when curators select IDM terms at different levels of the ontology, one being more general and the other more specific. For example, the interaction of the human proteins MDM2 and TP53 is listed in APID as (MDM2, TP53, 'anti tag Co-immunoprecipitation', PMID:17159902, INTACT:7156209) and also as (MDM2, TP53, 'affinity chromatography technology', PMID:17159902, BIOGRID:680279). Although biologists would naturally recognize one observation annotated twice at different granularities, the redundancy is not explicit. Implicit redundancy should not be confused with the common case where several experimental techniques are used in a single publication to validate a given PPI. Therefore, detecting implicit redundancies requires knowledge on IDMs.

Hereafter, we take advantage of the PSI-MI ontology to identify these two cases, as illustrated in Figure 1.

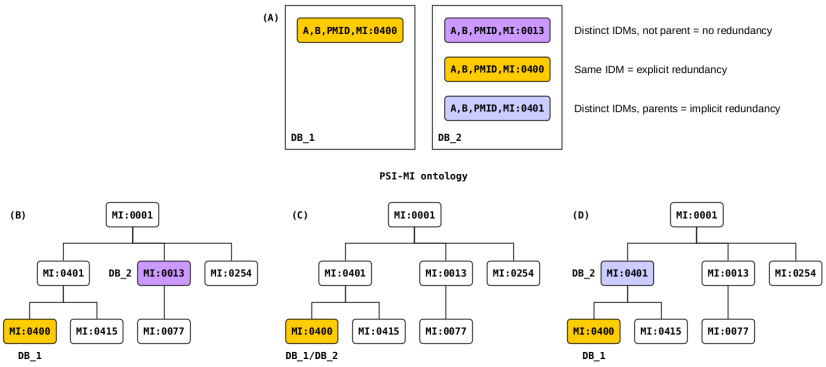

Fig. 1. Illustration of the different types of redundancy across primary databases. (A) Curation events from two databases (`DB_1` and `DB_2`). Depending on the IDM reported by `DB_2`, one can identify no redundancy (purple), explicit redundancy (yellow), or implicit redundancy (blue). Ontology representations of the different cases are presented in panels (B), (C) and (D).

2.2 Definitions

Following the notation introduced in 2.1, we consider two entries, E_i and E_j, of a meta-database, defined by their respective quintuplets of the form $(A, B, M_i, P_x, D_\alpha)$. Note that here we do not consider the experimental role of A and B, therefore all PPIs are symetric and the order of A and B is irrelevant.
E_i and E_j present explicit redundancy if and only if:

$$\begin{cases} E_i = & (A, B, M_i, P_x, D_a) \\ E_j = & (A, B, M_i, P_x, D_b) \end{cases}$$

E_i and E_j present implicit redundancy if and only if:

$$\begin{cases} E_i = & (A, B, M_i, P_x, D_a) \\ E_j = & (A, B, M_j, P_x, D_b) \\ M_i & \textit{is an ancestor of } M_j, \end{cases}$$

where an ancestor can be a direct or an indirect parent.

Ontologies such as PSI-MI (Sivade Dumousseau *et al.*, 2018) can be used to formalize the subsumption relations between IDMs. Note that with the notions provided in section 2.1, explicit and implicit redundancies might be observed among entries originating from different databases (inter-database redundancy, $D_a \neq D_b$) but also from the same database (intra-database redundancy, $D_a = D_b$). We will discuss later (Section 5.3) the meaning of intra-database redundancies, which can correspond either to multiple curation events, but also to variations of an IDM (for example, switching the experimental role (*'bait'* or *'prey'*) of the A and B proteins).

3 Methods

3.1 Source PPI datasets

PPI curation events integrated by APID were downloaded from the APID website on March 23, 2020, last update of APID in January, 2019) for two species (*Homo sapiens* and *Saccharomyces cerevisiae*) in the MITAB25 format (Kerrien *et al.*, 2007). These files aggregate the curated events from five primary databases in a standard format.

In MITAB25 formated data, each line represents a curation event. Interacting proteins are identified by their Uniprot accession numbers. The organism is identified with its NCBI taxonomy identifier. Various information on the experimental evidence is also provided, notably the PMID of the source publication and the PSI-MI code of the IDM used to detect the interaction. Some information such as the direction of the interaction (which protein was used as a *'bait'* and which as a *'prey'*) is not available in this format, but it is usually recorded in primary databases or in more recent MITAB formats (MITAB27). If necessary, missing information might be retrieved using the primary database interaction identifier which is provided and offers full tractability.

3.2 RDF schema and triplestore

The global RDF schema used to integrate all information is presented in Figure 2. It relies on the following ontologies:

- Biological Pathway Exchange (BioPAX) is an ontology developed as a standard for representing molecular interactions, including protein-protein interactions (Demir *et al.*, 2010). We followed the level 3 of the BioPAX specification.
- Proteomics Standards Initiative-Molecular Interactions (PSI-MI) is an ontology edited by the HUPO-PSI. It is dedicated to describe experimental IDMs (Sivade Dumousseau *et al.*, 2018). We used version 1.2.

Raw PPI curation events from the MITAB file were first imported into a MySQL database. A Perl script was used to connect to this database, to exclude curation events that are not considered by APID (see below), and to convert it into a RDF dataset following the BioPAX v3 standard. The resulting interaction data were merged with the PSI-MI ontology, available as an OWL file, into a triplestore powered by the Apache Foundation's JENA suite (v3.14.0). The complete workflow is described in Supplementary Figure S1.

In its integration process, the APID meta-database does not consider curation events annotated with IDMs that do not correspond to a specific experimental method (Alonso-López *et al.*, 2019). To be able to compare our results with APID, we also excluded from our analysis the very same curation events. These are the ones annotated with the IDMs 'molecular interaction', 'interaction detection method', 'biophysical', 'experimental interaction detection', 'inference', 'inferred by author', 'inferred by curator', 'in vitro', 'in vivo', 'unspecified method', or 'phenotype-based detection assay'.

3.3 SPARQL queries

Queries were run using SPARQL Protocol and RDF Query Language (SPARQL). The JENA suite was used to run the SPARQL queries. All queries used to detect redundancies are available in supplementary data

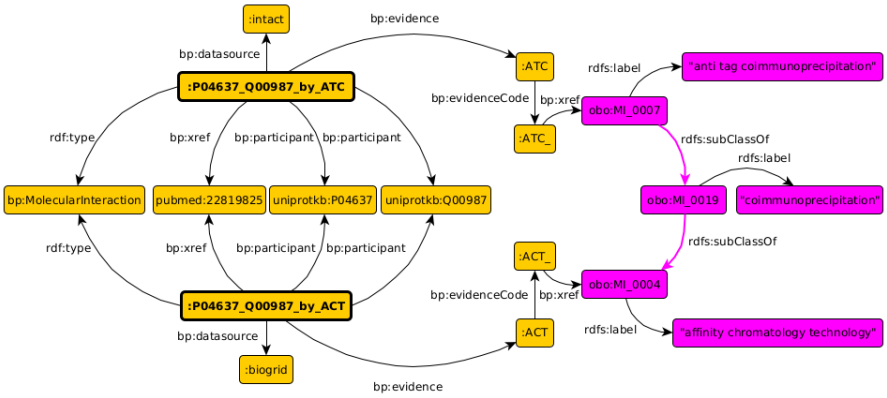

Fig. 2. Scheme representing two curation events reporting the interaction between the ubiquitin ligase MDM2 (UniprotKB: P04637) and the tumor protein 53 (UniprotKB: Q00987) in the BioPAX level 3 ontology (yellow nodes). These two curation events (highlighted in bold) were annotated by different databases (BioGRID and Intact). They refer to the same publication (PMID: 22819825), but different IDMs were used to annotate the interaction ('anti tag coimmunoprecipitation' and 'affinity chromatography technology'). The PSI-MI ontology (purple nodes) reveals that 'affinity chromatography technology' is an ancestor of 'anti tag coimmunoprecipitation', indicating an implicit redundancy between the two curation events.

(Figures S2, S3, S4, S5, S6, S7). As an example, Figure 3 presents the SPARQL query used to detect implicit redundancies in curation events, if one term is an ancestor of the other in the PSI-MI ontology. For each implicit redundancy detected, we conserved only the curation event with the most precise IDM.

```
SELECT DISTINCT ?p1 ?p2 ?pmid ?dm_name1
WHERE {
    ?ppi1 rdf:type bp:MolecularInteraction ;
          bp:participant ?p1, ?p2 ;
          bp:xref ?pmid ;
          bp:evidence ?dm_name1 .
    ?dm_name1 bp:evidenceCode ?m_vocab1 .
    ?m_vocab1 bp:xref ?dm_code1 .
    FILTER ( STR(?p1) < STR(?p2) )
    FILTER NOT EXISTS {
        ?ppi2 rdf:type bp:MolecularInteraction ;
              bp:participant ?p1, ?p2 ;
              bp:xref ?pmid ;
              bp:evidence ?dm_name2 .
        ?dm_name2 bp:evidenceCode ?m_vocab2 .
        ?m_vocab2 bp:xref ?dm_code2 .
        ?dm_code2 rdfs:subClassOf+ ?dm_code1 .
    }
}
```

Fig. 3. SPARQL query to select curation events without explicit nor implicit redundancies. (Note: prefixes are not shown)

3.4 Availability and implementation

The code is available at `https://gitlab.com/nnet56/reproducible-interactome`. The results are available at `https://reproducible-interactome.genouest.org/` and on the Zenodo open data repository (doi:10.5281/zenodo.5595037). The non-redundant interactomes are also accessible on the NDEx platform to facilitate their analysis and manipulation with classical algorithms (doi:10.18119/N94302 (human), doi:10.18119/N97S4D (yeast)).

4 Results

4.1 Overview of analyzed curation events

We analysed the same curation events as the APID database to assess the efficiency of redundancy detection methods. A summary of these curation events is presented in Table 1. The downloaded MITAB files contain $700,484$ curation events for *Homo sapiens* and $305,102$ for *Saccharomyces cerevisiae* (hereinafter referred to as human and yeast, respectively). Together, BioGRID and IntAct represent approximately 85% of all curation events in both species. The contribution of HPRD and BioPlex, restricted to human data, accounts for 13.9% of human curation events. For both species, most PPIs appear in only one or two curation events. PPIs reported by a single curation event represent 49.3% and 60.7% of interacting pairs in human and yeast, respectively.

4.2 Interaction detection methods (IDMs)

The most frequent IDMs in all curation events are listed in Table 1. Among them, 'affinity chromatography technology', 'tandem affinity purification', 'anti tag coimmunoprecipitation' and 'two hybrid' cover more than 58% of human and 76% of yeast curation events. Interestingly, these IDMs include terms with parent–child relationships in the PSI-MI ontology. For example, 'affinity chromatography technology' is a direct ancestor of 'anti tag coimmunoprecipitation'. The presence of such chains is suggestive of possible implicit redundancies between curation events, as defined in sections 2.1 and 2.2.

4.3 Quantification of implicit redundancies

Thanks to the expressiveness of the SPARQL language, we identified both explicit and implicit redundancies among curation events (example query in Figure 3). For constituting a non-redundant dataset, we selected the most precise curation events and discard the redundant and less precise ones since they do not add information.

The occurrence of redundancy among curation events is significant (Table 2). We detected and discarded $73,991$ (11.1%) and $40,266$ (13.7%) implicitly redundant curation events for human and yeast, respectively. Taking into account both explicit and implicit redundancies resulted in removing 30.9% of curation events for human and 35.4% for yeast.

Table 1. Human and yeast curation events (CEs) analysed in this study. Excluded Interaction Detection Methods (IDMs) concern 5.00% ($n = 35,000$) of all curation events in human and 3.87% ($n = 11,809$) in yeast. Only IDMs annotated with a frequency higher than 2% are shown.

Contributing Databases			**Most frequent Interaction Detection Methods**			**Currations events for** (P_a, P_b)		
Databases	**CEs**	**(%)**	**Interaction Detection Methods**	**Counts**	**(%)**	**Occurrences**	**Counts**	**(%)**
Human								
BioGRID	378,910	(54.1%)	Affinity chromatography technology	291,621	(41.63%)	One	161,031	(49.30%)
IntAct	215,577	(30.8%)	Two hybrid	71,969	(10.27%)	Two	91,742	(28.08%)
BIOPLEX!	55,151	(7.9%)	Anti tag coimmunoprecipitation	49,428	(7.06%)	[3-10[	69,015	(21.13%)
HPRD	42,327	(6.0%)	Pull down	42,423	(6.06%)	[10-50[	4,763	(1.46%)
DIP	8,519	(1.2%)	Biochemical	40,544	(5.79%)	≥ 50	113	(0.03%)
			Anti bait coimmunoprecipitation	27,745	(3.96%)			
			In vivo	21,118	(3.01%)			
			Two hybrid array	20,813	(2.97%)			
			Validated two hybrid	14,525	(2.07%)			
Yeast								
BioGRID	133,998	(43.9%)	Affinity chromatography technology	88,681	(29.07%)	One	83,799	(60.73%)
IntAct	130,025	(42.6%)	Tandem affinity purification	84,842	(27.81%)	Two	28,496	(20.65%)
DIP	41,079	(13.5%)	Anti tag coimmunoprecipitation	35,363	(11.59%)	[3-10[	21,792	(15.79%)
			Two hybrid	24,752	(8.11%)	[10-50[	3,799	(2.75%)
			Pull down	13,960	(4.58%)	≥ 50	99	(0.07%)
			Inferred by author	10,894	(3.57%)			
			Protein complementation assay	6,825	(2.24%)			
			Enzymatic study	6,817	(2.23%)			

Table 2. Impact of the removal of both explicit and implicit redundancies on the number of curation events and on the apparent size of the reproducible interactome, for human and yeast. (EEs: Experimental Evidences)

	Human	(%)	**Yeast**	(%)
Curation events				
Initial curation events	665,484	(100%)	293,293	(100%)
Curation events without explicit redundancies	534,140	(80.3%)	229,630	(78.3%)
Curation events without explicit and implicit redundancies	460,149	(69.1%)	189,364	(64.6%)
Apparent size of the reproducible interactome (PPIs supported by ≥ 2 EEs)				
Initial	159,192	(100%)	52,313	(100%)
Without explicit redundancies	111,009	(69,7%)	40,235	(76.9%)
Without explicit and implicit redundancies	70,554	(44,3%)	21,311	(40.7%)

Importantly, detection of redundancy between curation events has a strong impact on the apparent size of the reproducible interactome (i.e PPIs supported by at least two experimental evidences) (Table 2, Supplementary Figures S8 and S9). For human, the reproducible interactome drops from $159,192$ to $70,554$ PPIs (-55.7%: -30.3% due to explicit redundancies and -25.4% due to implicit ones). For yeast, the impact of redundancies is even worse, with a drop of the reproducible interactome from 52,313 PPIs to 21,311 after removal of both explicit and implicit redundancies (-59.3%: -23.1% due to explicit redundancies and -36.2% due to implicit ones). In other words, for human, discarding 11.1% of implicitly redundant curation events accounts for reducing by 25.4% the reproducible interactome. Similarly, for yeast, discarding 13.7% of implicitly redundant curation events accounts for reducing by 36.2% the reproducible interactome.

4.4 Implicit redundancies mostly result from the integration of the different primary databases

We then investigated whether implicit redundancy was already present in source databases (intra-database redundancy), or if it was a consequence of the integration of different source databases (inter-database redundancy). The vast majority originates from inter-database redundancies for both human (91.1%) and yeast (95.0%) (see Supplementary Tables S1 and S2). The couple of databases that generates the largest part of the implicit redundancies is BioGRID and IntAct. This is consistent with the fact that BioGRID and IntAct are the two most contributing source databases. Intra-database redundancies will be further discussed in section 5.3.

4.5 Frequently redundant identification methods

We computed the frequency of the pairs of detection methods involved in implicit redundancies. For human, the most frequent implicitly redundant couples of IDMs and their parent-child relationships in the PSI-MI ontology are displayed in Figure 4.

The most frequent couple is 'affinity chromatography technology' and 'anti tag coimmunoprecipition', which is responsible for $25,333$ redundancies. The term 'affinity chromatography technology' is also frequently observed with other descendants such as "pull down" ($n = 9,896$), 'anti bait coimmunoprecipitation' ($n = 6,617$), or "tandem affinity purification" ($n = 5,968$). Two-hydrid techniques are also introducing redundancies, for example with 'two hybrid', and its descendants 'two hybrid array' ($n = 16,113$), 'two hybrid prey polling approach' ($n = 11,238$), 'validated two hybrid' ($n = 11,123$), or 'two hybrid pooling approach' ($n = 10,713$). A similar situation is observed in yeast (the complete list of implicit redundancies for both human and yeast is available as Supplementary Tables S3 and S4). Implicit redundancies are thus widespread all along the PSI-MI ontology, and not limited to binary IDMs. This highlights the need for a general approach to reconcile

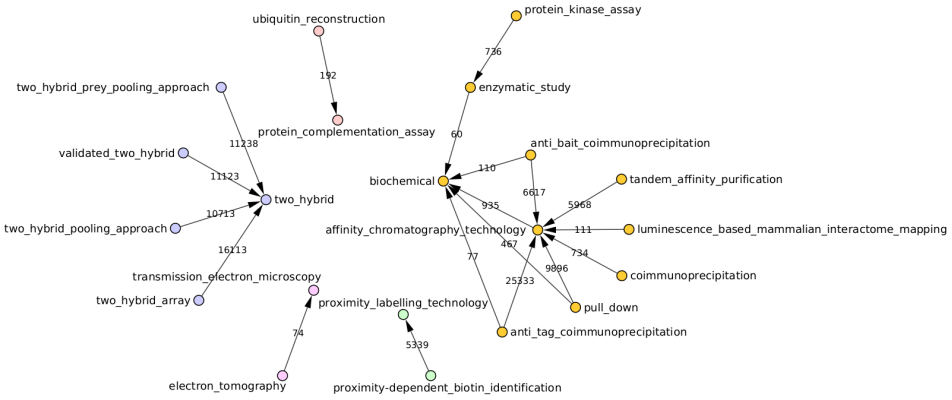

Fig. 4. Couples of related interaction detection methods (IDMs) from the PSI-MI ontology frequently identified in implicit redundancies of human PPIs. The arrows connect the most specific to the most general term according to the PSI-MI ontology. Only implicit redundancies with at least 50 occurrences are shown. Nodes connected to a common IDM are represented with the same color.

curation events during the integration of multiple primary databases.

The fact that implicit redundancies are observed between very different terms of the PSI-MI ontology suggests that different primary databases have different policies for annotating IDMs, as previously noted for IntAct and BioGRID (Alonso-López *et al.*, 2019). We therefore further analysed the IDMs used by each primary database.

We observed that IntAct and DIP use a wide range of IDMs for both human and yeast PPIs (165 for IntAct and 89 for DIP) while BioGRID, HPRD and BioPlex use much fewer (12, 3 and 1 IDMs, respectively) and more general IDMs. Hence, the strong discrepancies in database annotation policies are the source of inter-database implicit redundancies.

Overall, we observed that implicit redundancy (i) occurs between a wide range of the PSI-MI ontology terms, regardless of the species, (ii) mostly results from the integration of different primary databases with different annotation policies, and (iii) happens for all database combinations.

5 Discussion

The construction of a reliable interactome demands to combine interaction data produced by several independent experimental evidences and IDMs in order to reduce false positives. Since experimental evidences are curated and stored in several primary databases, a unification of these databases is required. The Human Proteome Organization Proteomics Standards Initiative (HUPO-PSI) developed the PSICQUIC specification and web services that facilitate data retrieval from multiple databases and assist their integration but do not elaborate on redundancy detection (del Toro *et al.*, 2013). In several (meta-) databases, PPIs are annotated with a confidence score, which is calculated using the number of independent experimental evidences and the nature of IDMs (Villaveces *et al.*, 2015). To be relevant, these algorithmic require reliable, non-redundant, datasets of experimental evidences. Therefore, several primary databases have decided to coordinate their curation efforts in the frame of the IMEx consortium in order to provide a single non-redundant set of homogeneously annotated protein interaction data (Orchard *et al.*, 2012; Porras *et al.*, 2020).

Here, we propose a formalisation of both explicit and implicit redundancy between experimental evidence entries in order to integrate PPIs from any database that uses the PSI-MI ontology. Knowledge about IDMs is extracted from the PSI-MI ontology, while the method to identify redundancies is based on Semantic Web technologies.

5.1 The Semantic Web is adapted for identifying implicit redundancies

Alonso-López *et al.* (2019) pointed two problems related to redundancy identification: (i) there may be a parent-child relationship between IDM terms, and (ii) the path from a child term to its ancestors may not be unique due to multiple inheritance. We propose the notion of **implicit redundancy** to address the logical implications of two database entries describing the interaction of the same protein pair with IDMs that have a descendant-ancestor relationship. The Semantic Web is designed to perform integrated reasoning on data annotations and ontologies. In particular, it makes handling simple and multiple hierarchies straightforward. In the raw data of APID that aggregates BioGRID, IntAct, HPRD, BioPlex and DIP, we were able to identify both explicit and implicit redundancies. Our work reveals that implicit redundancies are a widespread phenomenon resulting from the different curation choices of the various databases and that it is of similar importance than explicit redundancies. Therefore, we demonstrated the relevance of both the notion of implicit redundancy and of the choice of the Semantic Web as a technical framework for addressing the redundancy identification problem. Moreover, new explicit and implicit redundancies will continue to occur over the natural updates of the various databases.

The PSI-MI ontology that describes the IDMs is evolving. For example, during the time of our project, we noticed that the term 'three hybrid', which was initially a child of the term 'two hybrid', is now a child of 'transcriptional complementation assay'. This modification is highly relevant since 'two hybrid' is a binary identification method, whereas 'three hybrid' is not, and having a non-binary identification method as a direct child of a binary one was not consistent. Therefore, just like the databases are regularly updated, the ontologies are also corrected and enriched, which also has an incidence on redundancies. By allowing to automate redundancy detection as the integration of databases scales up, the Semantic Web facilitates the reliable interpretation of the results in the perspective of the construction of a reproducible interactome.

5.2 Widespread inter-databases implicit redundancies

Implicit redundancies primarily arise from the integration of different databases (91.1% and 95.0% of inter-database redundancies for human

and yeast, respectively). In our study, we clearly highlight that this is due to the granularity of IDMs used in the primary databases. Indeed, while some databases like IntAct refer to numerous detailed terms from the PSI-MI ontology (165 and 89 terms used to annotate human and yeast PPIs, respectively), other databases like BioGRID merely use general and high level terms (only 12 terms used for both human and yeast).

Therefore, if the integration of different PPI databases is necessary to better cover the interactome, a particular attention has to be paid to detect the widespread inter-database implicit redundancies. A simple method could be to define priorities between databases depending on whether they use precise or general terms to annotate PPIs. In case of multiple curations events referring to the same proteins and the same PMID, the ones from the database with the highest priority would be selected. However, this would be an approximate approach whereas we propose an exact solution, robust to possible changes of annotation policy by primary databases.

Primary databases of the IMEx consortium coordinate and share their curation efforts to produce a non-redundant dataset of PPI experimental evidences (Orchard *et al.*, 2012). IMEx members use common curation rules to harmonize their annotation process. The unicity of the curation events is ensured by allowing PPIs from a given PMID to be annotated only once, and all data are centralized in IntAct. Both this work from the IMEx consortium and ours emphasize the need for a general approach to assemble non-redundant PPI datasets.

5.3 Intra-database redundancies

Our analysis also identified a significant number of apparently redundant curation events within primary databases (Supplementary Figures S8 and S9). Such intra-database redundancy may originate from multiple independent annotations of identical experimental evidences within primary databases, as noted by Alonso-López *et al.* (2019). Yet, further inspection of such curation events indicates that intra-database redundancy primarily occurs when independent experiments from the same publication have been annotated in a given database with identical or related IDMs, leading to apparent explicit or implicit intra-database redundancies. For instance, we observed that the vast majority of the explicit intra-database redundancies originating from BioGRID are due to PPIs probed with both partners as baits and preys (6229 out of 8696 explicit redundancies involving exactly two curation events for yeast and 12283 out of 15385 for human). Intra-database redundancy can also occur when a PPI has been identified with a high-throughput experiment and then validated using the same or a related method performed at low-throughput. Hence, this currently leads to the unification of curation events that actually report distinct experimental evidences. To correct this, our method could be extended by taking into account additional information, such as the experimental role of each protein.

5.4 Towards a reproducible interactome

The size of the reproducible interactome is drastically impacted by removing redundancies for both human (-55.7%) and yeast (-59.3%), and we show that this is largely due to implicit redundancies. Indeed, we observe that filtering the curation events involved in implicit redundancy (11 to 14 %) leads to a drastic (25 to 36 %) reduction of the apparently reproducible interactome. This implies that a large number of PPIs currently considered as reproducible actually relies on integration artefacts. Thus, more experimental data are still needed to further improve the size and confidence level of the reproducible interactome. Information on PPIs that have not yet been reproduced can help to prioritize such experiments. Knowledge-based methods as presented in this article will be necessary to support the integration of the continuously increasing experimental evidences and publications.

Acknowledgements

The GenOuest platform provided computational support and Web hosting.

Funding

This work has been supported by Univ Rennes with a Defi Emergent 2019 grant to EB and GR.

References

Alonso-López, D. *et al.* (2016). APID interactomes: providing proteome-based interactomes with controlled quality for multiple species and derived networks. *Nucleic Acids Res*, **44**(W1), W529–535.

Alonso-López, D. *et al.* (2019). APID database: redefining protein–protein interaction experimental evidences and binary interactomes. *Database*, **2019**. baz005.

Bonetta, L. (2010). Interactome under construction. *Nature*, **468**(7325), 851–852.

Cafarelli, T. *et al.* (2017). Mapping, modeling, and characterization of protein–protein interactions on a proteomic scale. *Current Opinion in Structural Biology*, **44**, 201–210.

Calderone, A. *et al.* (2013). mentha: a resource for browsing integrated protein-interaction networks. *Nat Methods*, **10**(8), 690–691.

Calderone, A. *et al.* (2020). Using the MINT Database to Search Protein Interactions. *Curr Protoc Bioinformatics*, **69**(1), e93.

Das, J. and Yu, H. (2012). HINT: High-quality protein interactomes and their applications in understanding human disease. *BMC Syst Biol*, **6**, 92.

del Toro, N. *et al.* (2013). A new reference implementation of the PSICQUIC web service. *Nucleic Acids Res*, **41**(Web Server issue), W601–606.

Demir, E. *et al.* (2010). The BioPAX community standard for pathway data sharing. *Nature biotechnology*, **28**(9), 935–942.

Huttlin, E. L. *et al.* (2021). Dual proteome-scale networks reveal cell-specific remodeling of the human interactome. *Cell*, **184**(11), 3022–3040.

Kerrien, S. *et al.* (2007). Broadening the horizon–level 2.5 of the HUPO-PSI format for molecular interactions. *BMC Biol*, **5**, 44.

Kerrien, S. *et al.* (2012). The IntAct molecular interaction database in 2012. *Nucleic Acids Res*, **40**(Database issue), D841–846.

Keshava Prasad, T. S. *et al.* (2009). Human Protein Reference Database–2009 update. *Nucleic Acids Res*, **37**(Database issue), D767–772.

Klapa, M. I. *et al.* (2013). Reconstruction of the experimentally supported human protein interactome: what can we learn? *BMC systems biology*, **7**, 96.

Kotlyar, M. *et al.* (2019). IID 2018 update: context-specific physical protein-protein interactions in human, model organisms and domesticated species. *Nucleic Acids Res*, **47**(D1), D581–D589.

Luck, K. *et al.* (2020). A reference map of the human binary protein interactome. *Nature*, **580**(7803), 402–408.

Orchard, S. *et al.* (2012). Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. *Nat Methods*, **9**(4), 345–350.

Oughtred, R. *et al.* (2019). The BioGRID interaction database: 2019 update. *Nucleic Acids Res*, **47**(D1), D529–D541.

Porras, P. *et al.* (2020). Towards a unified open access dataset of molecular interactions. *Nature communications*, **11**(1), 6144.

Salwinski, L. *et al.* (2004). The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res*, **32**(Database issue), D449–451.

Sivade Dumousseau, M. *et al.* (2018). Encompassing new use cases - level 3.0 of the HUPO-PSI format for molecular interactions. *BMC Bioinformatics*, **19**(1), 134.

Szklarczyk, D. *et al.* (2019). STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res*, **47**(D1), D607–D613.

Turinsky, A. L. *et al.* (2010). Literature curation of protein interactions: measuring agreement across major public databases. *Database*, **2010**. baq026.

Villaveces, J. M. *et al.* (2015). Merging and scoring molecular interactions utilising existing community standards: tools, use-cases and a case study. *Database*, **2015**. bau131.

COLLEGE ECOLOGIE
DOCTORAL GEOSCIENCES
BRETAGNE AGRONOMIE ALIMENTATION

L'INSTITUT agro Rennes Angers

Titre : Intégration et analyse de données biologiques hétérogènes par exploitation de graphes multicouches pour mieux comprendre les variations d'efficience alimentaire chez le porc

Mot clés : Efficience alimentaire, Graphe multicouche, Intégration de données, Multi-omiques, Web sémantique

Résumé : Les progrès technologiques d'étude du vivant ont conduit à une explosion de données multimodales et multicentriques. Ce phénomène soulève de nombreuses questions liées au stockage, à la standardisation et à l'analyse de ces données massives. Ainsi, ce travail de thèse porte sur le développement d'une méthode intégrative d'analyse de données biologiques, pour en extraire de la connaissance. Pour prendre en compte leur forte interdépendance, cette approche consiste à intégrer différents types d'entités biologiques (ARNm, protéines, métabolites, caractères observables) qui sont habituellement étudiés indépendamment les uns des autres. La solution informatique élaborée permet d'intégrer ces données hétérogènes dans un graphe multicouche, avec une couche par type d'entités. L'originalité est de relier les éléments d'une couche ou de couches différentes par des propriétés extraites des bases de données et de connaissances publiques à l'aide de technologies du Web Sémantique. A partir de ce graphe, le but est de caractériser les relations entre un groupe de molécules d'intérêt grâce à des métriques de la théorie des graphes. La méthode développée a été appliquée à des jeux de données expérimentaux (transcriptomique, métabolomique et phénotypes animaux) pour décrire et comprendre les relations entre les molécules et leur importance dans la variation d'efficience alimentaire de porcs. L'efficience alimentaire est un phénotype clé pour contribuer à un élevage durable, mais complexe. Ce travail a permis de mettre à disposition des méthodes d'analyse novatrices, à différentes échelles de l'organisation du vivant, favorisant une meilleure compréhension des processus biologiques.

Title: Integration and analysis of heterogeneous biological data through multilayer graph exploitation to gain deeper insights into feed efficiency variations in growing pig

Keywords: Data integration, Feed efficiency, Multilayer graph, Multi-omics, Web Semantic

Abstract: Recent technological advancements in biological data acquisition have resulted in an explosion of multimodal and multicentric data. This phenomenon raises numerous questions regarding the storage, standardization, and analysis of these massive datasets. This thesis focuses on the development of an integrative method for analyzing biological data to extract knowledge from them. To account for their strong interdependencies, this approach involves integrating different types of biological entities (mRNA, proteins, metabolites, observable traits) that are typically studied independently. The devised computational solution enables the integration of these heterogeneous data into a multilayer graph, with each layer representing a specific type of entity. The novelty lies in linking elements within a layer or across different layers by utilizing properties extracted from public knowledge databases through Semantic Web technologies. Based on this graph, the objective is to characterize the relationships among a group of molecules of interest using graph theory metrics. The method was applied to experimental datasets (transcriptomics, metabolomics and animal phenotypes) to describe and understand the relationships between specific molecules and determine their importance in feed efficiency variations in growing pigs. Feed efficiency is a key phenotype for sustainable farming, but is recognized as complex. This work provides innovative analysis methods to analyze and integrate various levels of biological organization, facilitating a better understanding of biological processes.