Identification de signature causale pathologie par intégration de données multi-omiques

Méline Wery

UNIVERSITÉ DE RENNES 1

THÈSE DE DOCTORAT DE

L'UNIVERSITÉ DE RENNES 1

ÉCOLE DOCTORALE N° 601
Mathématiques et Sciences et Technologies de l'Information et de la Communication
Spécialité : *Informatique*

Par
Méline WERY

Identification de signature causale pathologique par intégration de données multi-omiques

Thèse présentée et soutenue à l'IRISA, Rennes, le 16 décembre 2020
Unité de recherche :
Équipe Dyliss, Univ Rennes, Inria, CNRS, IRISA
Plateforme de Sciences Translationnelles, SANOFI R&D, Chilly-Mazarin

Rapportrices avant soutenance :

Laurence CALZONE	Ingénieure de Recherche à l'Institut Curie, Paris
Fleur MOUGIN	Maîtresse de conférence à l'Université de Bordeaux

Composition du Jury :

Président :	Franck DELAPLACE	Professeur à l'Université de Paris-Saclay - Evry
Examinateurs :	Franck DELAPLACE	Professeur à l'Université de Paris-Saclay - Evry
	Vassili SOUMELIS	PU-PH à l'hôpital Saint Louis, Paris
Dir. de thèse :	Olivier DAMERON	Professeur à l'Université de Rennes 1
Co-dir. de thèse :	Anne SIEGEL	Directrice de Recherche CNRS à Irisa Rennes
Encadr. de thèse :	Emmanuelle BECKER	Maîtresse de conférence à l'Université de Rennes 1
	Charles BETTEMBOURG	Data Scientist à SANOFI R&D, Chilly-Mazarin
Invité :	Emmanuel OGER	PU-PH à l'Université de Rennes 1

"Il y a des moments où tout réussit.
Il ne faut pas s'effrayer, ça passe."

Jules Renard

Remerciement

Je tiens tout d'abord à remercier Laurence Calzone, Fleur Mougin, Franck Delaplace, Vassili Soumelis et Emmanuel Oger d'avoir accepté de faire partie de mon jury.

Que serait une thèse sans encadrant... Un grand merci général à mes deux directeurs de thèse, Anne Siegel et Olivier Dameron ainsi qu'à mes deux autres encadrants, Emmanuelle Becker et Charles Bettembourg. Votre disponibilité et votre écoute dans les bons moments et ceux de doutes m'ont permis d'arriver au bout de cette aventure. Merci pour votre soutien et vos encouragements même quand la technologie ne suivait pas ! Je garde de cette thèse un très bon souvenir grâce à vous.
Anne, merci de m'avoir accueillie dans l'équipe Dyliss dès mon stage de M2. Je retiens de toujours chercher l'effet "Wahou" dans les histoires scientifiques que l'on raconte !
Olivier, merci pour ton soutien askomien tout au long de cette thèse, les fous rires en réunion ou lors des pauses. J'ai hâte de voir toutes ces belles futures publications !
Emmanuelle, merci pour toutes ces conversations sur l'avenir à plus ou moins long terme, la recherche et le reste !
Charles, merci de m'avoir montré d'autres opportunités de carrière, de ta disponibilité et des nombreuses conversations qu'on a pu avoir sur autres choses que la thèse !

Merci aussi aux équipes qui m'ont accueilli et supporter (!) pendant ces 3 ans. Aux 3 équipes symbiotiques : l'équipe Dyliss, mon équipe de coeur, Genscale et Genouest ! Merci pour ces superbes ambiances pendant les pauses et les séminaires. Les prochaines équipes où je serais, auront un gros challenge à relever à ce niveau-là. Merci aussi aux anciens symbiotes, Lucas, Clémence, Marie et tous les autres qui ont gentiment accepté de répondre à mes nombreuses interrogations techniques et scientifiques quand ils étaient présents. Á Xavier dont mon harcèlement permanent n'a pas fait trop fuir ! Á mon ancien co-bureau, Efflam qui m'a de nombreuses fois aidé sur la partie *admin sys* car "je suis qu'une biologiste moi, tout ces trucs techniques j'y connais rien". Et à Arnaud pour m'avoir fait découvrir Pandas, ce qui a littéralement changé ma vie de doctorante bio-informatique ! Enfin, celle qui est notre Blanche-Neige à tous, je te remercie Marie pour avoir magnifiquement géré tous mes déplacements sanofiens, les choix d'hôtels dont certains sont devenus des drôles d'anecdotes et toute l'administratif que je ne maîtrise pas toujours très

bien. A toute l'équipe de bioinformatique de SANOFI. Merci de m'avoir accueilli parmi vous, de m'avoir appris la vie dans le privé.

Tout au long de cette aventure, je n'ai pas été seule. Nous étions toute une équipe de doctorants, (dont les 7 merveilles du monde ou les 7 nains de Disney, c'est comme on veut), à connaître des hauts et des bas, à se soutenir chacun notre tour. Mais surtout à avoir des conversations plus ou moins philosophiques, ce qui nous permettait de nous évader le temps d'un café, d'une pause, d'un repas ou d'une soirée. Merci à vous, d'avoir été présents, de m'avoir laissé raconter ma vie et tout ce qui me passait par la tête ! La suite ne sera pas la même sans vous...
Aux petits derniers arrivés, bon courage pour ces 3 ans et n'oubliez la tradition de participer à Sciences en cour[t]s, c'est un passage obligé !

Aux un peu plus anciens Clara, Téo, Olivier, Pierre, Méziane, Arnaud, Grégoire, Nicolas et Kevin (*oppa* !), j'espère qu'on ne vous a pas tous fait peur avec nos moments de doutes ! Malgré tout, ça reste une superbe expérience et ce fut un plaisir de travailler avec vous et de discuter de beaucoup, beaucoup de choses essentielles (et oui Disney et les dramas sont des essentiels de vie !)
Enfin, les vieux comme moi je vous remercie encore une fois (c'est mon côté généreux) de tous ces moments passés ensemble et gravés ...

Lucas, même si tu es le plus vieux de nous puisque tu es déjà docteur, merci pour ta bonne humeur, tes connaissances illimitées et ta capacité de compréhension orale qui t'es propre.

Maël, tu resteras malgré toi, associé à Obélix et au fameux cotonnier, cet animal méconnu !

Hugo, je garderais ton super secret de vampire de Twilight avec moi ;D !

Lolita, j'ai adoré parler avec toi, le soir après le boulot ou parce qu'on avait besoin de faire une pause. Discuter de l'avenir (éleveuse de lapins peut toujours fonctionner !), de nos possibilités et de nos envies mais aussi de nos habitudes de vie ! On se donne rdv en janvier pour nos nouvelles résolutions.

Marine, j'ai adoré discuter avec toi de drag queen, de manga, de drama mais aussi de Web Sémantique, ce qui nous a relié pendant ces 3 ans. J'ai été plus que ravie d'être ta partenaire en tant que réalisatrices de courts métrages (Wery et Louarn^2 !) mais aussi en tant que sportives de haut niveau !

Et enfin, *last but not least*, Wesley mon double différent. Qu'aurait été cette thèse sans toi... Tout ces délires et ces nombreuses discussions, parfois pour ne rien dire, juste déconcentrer l'autre. Et que dire de cette discrétion qui nous caractérise tellement bien ! Ton égo prendra toujours une place importante dans ma vie ! Une ultime fois, merci.

A tout ce qui ont croisé mon chemin et qui m'ont beaucoup apporté mais que j'ai oublié au moment d'écrire ces remerciements.

Aux membres de l'association Nicomaque, merci de m'avoir comptée parmi vous pendant

ces 2 dernières années et de m'avoir permis de participer aux aventures que sont Sciences en cour[t]s et le Forum Doctorant-Entreprise.

Il n'y a pas que le travail dans la vie, il y a aussi tout ce qui se passe à côté. A tous mes amis, qui m'ont soutenu pendant toutes ces années. A ceux qui sont géographiquement les plus proches mais aussi ceux que je ne pouvais pas voir plus souvent. Merci à vous.

La bande du week-end, Marina, Jérémy, Manu et Lucie, vous m'avez permis de m'évader quand je rentrais dans ma région. Merci pour toute cette bonne ambiance !

Méryl, je ne désespère pas de venir te voir au Panama et qu'on se raconte nos vies. On s'est connu au tout tout début de ces études qui se sont avérées très longues, mais le lien perdure indéfiniment. Merci à toi ma petite coccinelle.

Sophie, tu m'as montré la voie de la bio-informatique et m'as conseillée dans mon choix de réorientation. Merci pour ton aide dans ma vie parisienne. A très vite j'espère.

Patricia, pouvoir discuter avec toi de tout et de rien, de mon choix de carrière et du tien. On finira peut être par travailler ensemble qui sait. Merci d'être présente.

Et enfin, à toi, ma BFF, ma soeur de coeur, mon âme soeur. On a commencé toute cette aventure de recherche ensemble, à coup de pipettage sur le son de Wham !. Toi qui est déjà docteur et qui m'a soutenue durant ces périodes compliquées où tu étais déjà passée. Tu es toujours là même à distance. Un énorme merci Claudie.

Enfin mes derniers remerciements vont à ceux qui me soutiennent depuis toujours dans tous mes choix.

A ma famille, qui me demande toujours quand je vais finir mes études. Je peux désormais vous le dire, c'est bon je suis arrivée au bout !

A ma belle-famille et à ma belle-soeur, merci d'avoir été là pour me permettaient de me vider la tête.

A ma grand-mère qui me demande toujours de lui expliquer ce que je fais et qui écoute très attentivement quand on en parle à la TV. Merci pour ce soutien indéfectible.

A mes parents et mon frère, vous qui m'avez toujours encouragée à faire ce que j'aimais, m'avez soutenue pendant toutes ces longues années et qui étiez là dans mes moments de doute. Merci pour votre écoute et votre présence. Je ne vous remercierais jamais assez pour tout ça. Ce travail est aussi un peu le votre.

Enfin, à toi Nicolas. Tu as connu tous mes choix d'études, depuis médecine jusqu'à cette dernière étape qu'est la thèse. Ce fut long mais je te remercie pour ta patience, ton soutien et ton amour. J'écris cette dernière page qui se tourne et je suis prête pour une nouvelle qui s'ouvre

pour nous deux...

SOMMAIRE

Liste des acronymes

ADN	Acide DesoxyRibonucléique
BILAG	British Isles Lupus Assessment Group
CNV	Copy Number Variation
DEG	Differentially Expressed Gene
eICTL	expression Individually-Consistent Trait Loci
eQTL	expression Quantitative Trait Loci
FCA	Formal Concept Analysis
GRN	Gene Regulatory Network
IFN	Interferon
IMI	Innovative Medicines Initiative
KEGG	Kyoto Encyclopedia of Genes and Genomes
LTh	Lymphocyte T helper
MAI	Maladie AutoImmune
MCA	Multiple Correspondance Analysis
PKN	Prior Knowledge Network
RDF	resource Description Framework
SLE	Systemic Lupus Erythematosus
SLEDAI	Systemic Lupus Erythematosus Disease Activity Index
SNP	Single Nucleotide Polymorphism
SPARQL	SPARQL Protocol And RDF Query Language
TCGA	The Cancer Genome Atlas
URI	Uniform Resource Identifier
VCF	Variant Call Format

TABLE DES FIGURES

Liste des tableaux

Liste des publications

Article déjà publié :

Formalizing and enriching phenotype signatures using Boolean networks
***Méline Wery**, Olivier Dameron, Jacques Nicolas, Elisabeth Remy, Anne Siegel*
Journal of Theoretical Biology,Volume 467, 2019, https://doi.org/10.1016/j.jtbi.2019.01.015
Cet article est présenté dans le chapitre 4.

Article en cours d'écriture :

Identification d'eICTL par intégration de données multi-omiques
***Méline Wery**, Franck Auge, Charles Bettembourg,Olivier Dameron, Anne Siegel, Emmanuelle Becker*
Cet article traitera de l'apport de la thèse discuté dans les chapitres 2 et 3 (modèle d'intégration et recherche d'eICTL).

Optimisation de la requête SPARQL par la division en subquery
***Méline Wery**, Franck Auge, Charles Bettembourg, Emmanuelle Becker, Anne Siegel, Olivier Dameron*
Cet article méthodologique traitera de l'optimisation de la requête SPARQL discutée dans le chapitre 3.

INTRODUCTION

I Contexte biologique général : Les maladies auto-immunes

I.i Généralité sur les maladies auto-immunes

Le système immunitaire a pour rôle de protéger l'organisme contre les agressions extérieures comme les bactéries et les virus. Lorsque cette protection se retourne contre les constituants du « soi », on parle de maladies auto-immunes (MAI).

D'après le Centre National de Référence (CRMR) des maladies auto-immunes de Strasbourg [1], les maladies auto-immunes touchent actuellement près de 10% de la population mondiale. Depuis plusieurs années, près de 80 maladies sont recensées et ce nombre ne semble plus évoluer. Elles sont classées en deux groupes :

- les maladies spécifiques à un seul organe comme le diabète de type 1 ou la sclérose en plaque
- les maladies non spécifiques, qui touchent plusieurs organes, aussi appelées maladie systémique comme la polyarthrite rhumatoïde ou le lupus

Plusieurs MAI apparaissent préférentiellement chez les femmes [1]. Plusieurs hypothèses sont avancées pour expliquer ce phénomène, parmi lesquelles l'influence des hormones ou le rôle du chromosome X [2]. La majorité de ces maladies n'ont pas encore d'origine connue mais plusieurs facteurs semblent influencer la maladie. On parle alors de maladie multi-factorielle.

I.ii Origine du dysfonctionnement

Facteurs génétiques et environnementaux Plusieurs facteurs semblent impliqués dans le déséquilibre de la réponse immunitaire (Figure 1).

Parmi les facteurs soupçonnés d'avoir un impact sur le développement de la maladie, on trouve la prédisposition génétique [4]. En effet, il existe plusieurs formes familiales de certaines MAI. Généralement, les MAI résultent d'une accumulation de polymorphismes de gènes dont les conséquences, pris indépendamment, sont faibles sur le développement de la pathologie. C'est le cas par exemple des gènes HLA, qui codent pour des protéines dont le rôle est de distinguer les éléments de l'organisme des pathogènes [5].

Des facteurs environnementaux sont aussi à prendre en compte dans le déclenchement de la pathologie. Plusieurs pistes sont d'ailleurs envisagées comme les hormones féminines qui sont

1. https ://maladie-autoimmune.fr/

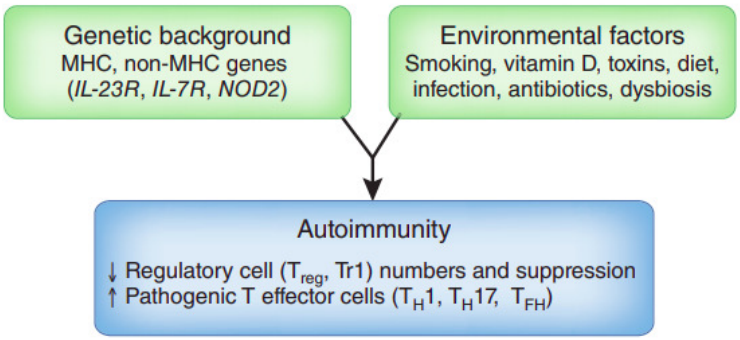


FIGURE 1 – **Schéma de l'influence des facteurs génétiques et environnementaux sur l'auto-immunité** (issu de [3]).

importantes dans certaines MAI, certains traitements anti-cancéreux qui peuvent induire la production d'auto-anticorps [6] ou encore le microbiote intestinal. En effet, des données épidémiologiques montrent une association entre ce microbiote et la survenue d'une maladie auto-immune [7]. Les modes de vie, tel que le régime alimentaire ou le fait de fumer, vont aussi favoriser le dérèglement du système immunitaire [3].

Si les causes ne sont pas encore complètement établies, les dérégulations cellulaires impliquées dans les MAI sont mieux définies.

Dérégulation du système immunitaire D'un point de vue biologique, la première étape dans l'apparition des MAI est la rupture de la tolérance du soi. La réponse auto-immune est ensuite médiée selon deux processus [8] (Figure 2).

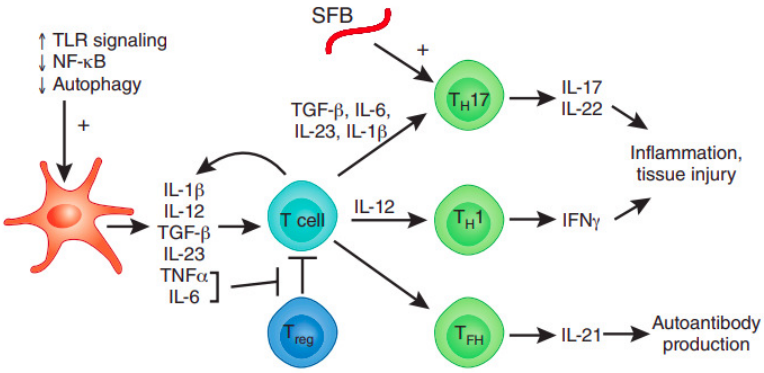


FIGURE 2 – **Schéma de la dérégulation cellulaire impliquée dans l'apparition de l'auto-immunité** (issu de [3]).

Un premier processus est celui des lymphocytes B. En effet, la reconnaissance des éléments du soi comme étant des agents pathogènes entraîne la production d'anticorps dirigés contre l'organisme, les auto-anticorps. La présence d'auto-anticorps fait partie des éléments analysés

pour le diagnostic de certaines MAI. Il en existe plusieurs selon la localisation de l'antigène, l'anti-corps et ne sont présents dans différentes MAI [9] : les anti-corps anti-nucléaires (ANA), les anticorps anti-cytoplasme de polynucléaires neutrophiles (ANCA), les anticorps dirigés contre le fragment constant (Fc) des immunoglobulines de type G, les anticorps anti-phospholipides et les anticorps anti-tissus.

Le second concerne les lymphocytes T autoréactifs vont jouer un rôle dans la lyse des cellules reconnues comme pathogènes et donc l'inflammation. Cette destruction se fait par la sécrétion de cytokines (interleukine et interféron) qui vont faire venir d'autres cellules immunitaires par chimiotactisme.

Si les mêmes dérèglements des mécanismes immunitaires semblent apparaître dans les MAI, les symptômes cliniques résultants de ces dysfonctionnements sont assez hétérogènes, en particulier pour les MAI systémiques.

I.iii Des symptômes hétérogènes et une stratification difficile

Les MAI systémiques sont des maladies auto-immunes qui touchent l'ensemble du système, impliquant plusieurs organes. Ces pathologies se manifestent cliniquement par plusieurs symptômes, généralement sous forme de crises. Les symptômes ne sont pas forcément les mêmes pour les patients diagnostiqués avec la même maladie et des pathologies différentes peuvent partager certains critères cliniques. Connaître le stade de la pathologie est important lors du diagnostic et se base sur plusieurs marqueurs selon les pathologies comme la mesure des auto-anticorps [10], le nombre et la sévérité de l'atteinte des organes. Si le stade est agressif, la prise en charge du patient sera différente (hospitalisation par exemple) et les traitements ne seront pas les mêmes. Cependant, l'hétérogénéité de l'aspect clinique entre les patients ne facilite pas l'obtention d'une stratification claire [11].

En 2014, un consortium européen, PRECISESADS [2], a mis en place une étude afin d'établir une nouvelle classification des patients, basée sur les mécanismes cellulaires et moléculaires communs plutôt que sur leurs symptômes seuls. Pour cela, 2 500 individus ont été recrutés, dont des patients de MAI systémiques variées. Ce projet vise à utiliser les données omiques (génomique, transcriptomique, épigénomique, métabolomique et protéomique) afin d'identifier des marqueurs et des mécanismes moléculaires qui caractériseraient mieux les patients. Une nouvelle stratification serait générée à partir de ces signatures pathologiques et permettrait, *in fine*, de mieux traiter les patients.

Tout l'enjeu de cette thèse est de définir une nouvelle méthode afin d'identifier de nouveaux marqueurs biologiques pour les patients atteints de MAI. Cette méthode prendra en compte l'aspect complexe et hétérogène de ce type de maladie en utilisant les ca-

2. https ://www.imi.europa.eu/projects-results/project-factsheets/precisesads

ractéristiques biologiques de chaque patient et non la population malade dans son ensemble. La recherche de ces marqueurs va aussi reposer sur la combinaison de plusieurs données omiques. Les données de génomique peuvent contenir des informations causales de la pathologie et les données de transcriptomique représentent l'expression des gènes. Pour cela, la méthode va s'appuyer sur des outils informatiques (web sémantique) et bio-informatiques (biologie des systèmes).

II Enjeux bioinformatiques : Traitement de données massives et hétérogènes, recherche de causalité

Les études cliniques ont pour principal but de répondre à une question biologique, médicale pour une pathologie. Par exemple, connaître l'effet d'un traitement ou identifier des biomarqueurs dans une maladie. Cependant, ces études nécessitent de filtrer les personnes recrutées par des critères d'inclusion et d'exclusion. La stringence de ces critères va influencer le nombre d'individus dans l'échantillon. De plus, avec l'essor des technologies à haut débit, la quantité et la nature des données générées ne fait qu'augmenter.

L'ensemble de ces facteurs explique que les études cliniques contiennent des données massives et hétérogènes [12]. Des analyses classiques en biologie comme des mesures statistiques, peuvent être compliquées à appliquer sur ce type de données. En revanche, plusieurs outils bioinformatiques peuvent aider à pallier ces deux problématiques.

II.i La problématique des données hétérogènes

Les technologies à haut débit vont générer plusieurs types de données selon le niveau moléculaire. Ces dernières peuvent différer dans leur format et leur contenu. La majorité des analyses comme l'expression de gènes ou l'abondance de protéines, métabolites traitent des données continues. Cependant, les données de génotypage des données discrètes et des données cliniques comme les scanners/IRM ou les radios sont sous forme d'images. Elles ont toutes un intérêt informatif sur la problématique biologique. Cependant, les analyses standards en biologie sont des mesures statistiques, qui se limitent généralement à l'utilisation de données continues.

Un autre problème inhérent à la biologie est la connexion entre ces différentes données [13]. Chaque couche omique va représenter un niveau biologique différent (ADN, gène, protéine, cellule, tissu, phénotype). Cependant, il existe des régulations entre ces couches plus ou moins directes. Il est nécessaire de prendre en compte ce facteur dans les analyses afin d'identifier les mécanismes moléculaires sous-jacents. C'est le cas par exemple des eQTLs qui sont des corrélations entre la présence d'une mutation (SNP) avec une variation d'expression de gènes et qui sont identifiés en comparant deux populations.

II.ii La problématique des données massives

En biologie, l'analyse principale demeure l'approche statistique. L'évolution de la technologie voit émerger l'analyse par apprentissage automatique (*machine learning*) des données omiques. Dans les deux cas, un point est crucial dans le choix des données. Il est nécessaire d'avoir un nombre d'individus important pour que la puissance statistique soit suffisante. Généralement, le nombre d'individus (n) doit même être supérieur au nombre de variables explicatives (p) (n»p). C'est ce sur quoi repose le concept de "Big Data", qui sont des données trop importantes pour être traitées par des approches classiques [14], [15].

Les études cliniques génèrent l'effet inverse, c'est-à-dire que la quantité de variables explicatives est bien supérieure au nombre d'individus (p»n). Par des approches statistiques, ce phénomène va entraîner une augmentation du taux de faux positifs [16]. En effet, il n'existe pas assez de contraintes dans la sélection des variables explicatives.

Comme expliqué précédemment, l'objectif de cette thèse est d'identifier des marqueurs relatifs au SLE. Cette recherche tente de répondre aux besoins de mettre en évidence des éléments causaux de la pathologie, pouvant expliquer des mécanismes moléculaires. Pour cela, deux approches sont développées et s'appuient sur des principes d'informatique et de bioinformatique. La première se base sur des données omiques hétérogènes issues d'une étude clinique et propose d'intégrer ces données en une structure de base de données. La recherche de biomarqueurs se fait via une analyse par raisonnement. La seconde utilise les connaissances des interactions entre les entités biologiques et se sert des dépendances de régulation pour identifier de nouveaux marqueurs.

II.iii L'intégration des données par les technologies du Web sémantique

L'intégration des données est nécessaire pour pouvoir exploiter au mieux l'ensemble des liens existants entre elles [17], [18]. Afin d'obtenir le maximum d'information, il est important que cette intégration passe par une structure bien définie. D'une part, l'annotation des entités et des interactions entre elles est un aspect important pour une bonne caractérisation des relations entre les couches omiques. D'autre part, les éléments mesurés dans une couche omique doivent pouvoir être retrouvés dans d'autres couches omiques. Cela n'est pas forcément le cas au vu des nombreux identifiants que peut prendre un même gène dans les bases de données.

L'exploitation de cette structure permet d'extraire les informations nécessaires. Dans le cadre cette thèse, l'analyse des données omiques est une analyse par raisonnement. Autrement dit, elle va exploiter les relations entre les entités ainsi que les différentes valeurs des entités pour identifier des biomarqueurs. Il reste important de pouvoir comparer les résultats obtenus avec d'autres bases de données fonctionnelles comme KEGG, Reactome.

Les technologies du Web Sémantique vont répondre à ces différents besoins de structuration et d'exploration des données intégrées. Le Web Sémantique est défini à la base pour structurer les informations contenues sur le Web afin de les relier [19]. Plusieurs représentations des données sont possibles et définies par le W3C. Mais toutes se basent sur l'utilisation d'URI (*Uniform Resource Identifier*) pour désigner les entités dans les données, aussi appelés ressources. Autrement dit, chaque entité est associée à un URI unique.

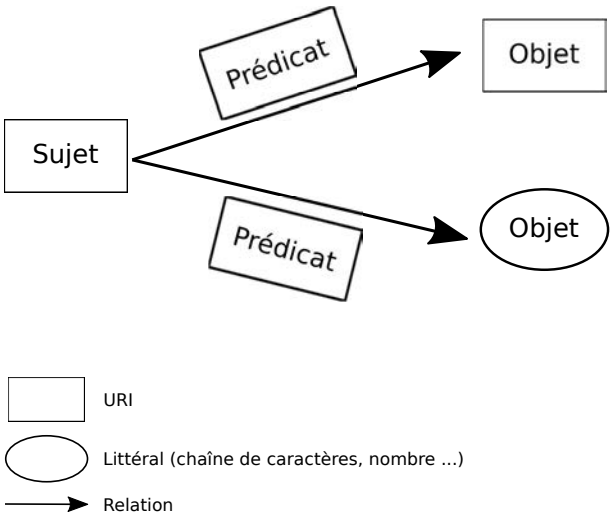


FIGURE 3 – **Représentation de la structure des triplets en RDF.** Le sujet et le prédicat sont des URI (*Uniform Resource Identifier*). L'objet peut être un URI ou une valeur littérale (chaîne de caractères, nombre etc ...).

Dans le cadre de cette thèse, c'est la représentation en RDF (Resource Description Framework [3]) qui a été privilégiée. Elle permet de représenter les relations entre toutes les entités et leurs valeurs sous forme de triplets *sujet-prédicat-objet* (Figure 3). Le sujet et le prédicat sont obligatoirement des URIs. L'objet peut être un URI si le triplet caractérise la relation (prédicat) entre une première entité (sujet) et une seconde (objet). Sinon l'objet est une chaîne de caractère, une valeur numérique etc...

Le triplet est un trio d'éléments orienté (sujet → prédicat → objet). Lorsque l'on souhaite visualiser toutes les connexions entre les triplets, on obtient un graphe orienté qui représente l'ensemble des relations entre les entités (graphe des données). C'est grâce à ce schéma que l'analyse par raisonnement est favorisée.

L'ensemble des triplets est stocké dans une base de données appelée *triplestore*. Le langage SPARQL va permettre d'interroger ces triplestores. L'analyse par raisonnement sur les données va passer par la construction de requêtes SPARQL en se basant sur le schéma RDF. De nombreuses bases de données biologiques sont représentées en RDF (KEGG, Reactome).

3. https ://www.w3.org/RDF/

L'élaboration de requêtes fédérées va combiner les résultats de l'analyse par raisonnement aux connaissances contenues dans ces bases.

Ainsi, le travail de recherche de cette thèse va proposer un cadre basé sur les technologies du Web Sémantique et permettre l'intégration et l'interrogation de données biologiques massives et hétérogènes.

II.iv La recherche de causalité par la biologie des systèmes

Un domaine de la bioinformatique a aussi pour objectif la structuration des données. Il s'agit de la biologie des systèmes qui va relier directement les éléments biologiques entre eux et non uniquement les couches omiques.

Les bases de données comme Reactome [4], Kegg [5] ou PathwayCommons [6] recensent de nombreuses connaissances sur les interactions moléculaires. Toutes ces connaissances sont issues d'expériences biologiques combinées aux expertises dans les différents domaines de la biologie. Chaque interaction est définie par la relation connue et dirigée entre deux entités. L'ensemble de ces interactions vont constituer des réseaux que l'on peut exploiter. Un réseau est défini comme étant un graphe dirigé dont les noeuds sont des éléments biologiques (gènes, protéines, métabolites) et les arcs correspondent aux relations typées entre ces entités.

Selon la classe des entités, plusieurs types de réseaux sont identifiés. Lorsque les entités sont des gènes et/ou des facteurs de transcription, ce sont des réseaux de régulation de gènes (*Gene Regulatory Network* ou GRN). Lorsque les noeuds du réseau sont des protéines, deux types de réseaux peuvent correspondre : d'une part des réseaux d'interaction protéine-protéine, qui caractérisent la formation de complexes et d'autre part, des réseaux de signalisation, montrant par exemple l'activation ou l'inactivation de protéines et par extension, l'implication de voies de signalisation. Enfin, lorsque les éléments correspondent à un ensemble de métabolites et d'enzymes, on parle de réseaux métaboliques. Ils permettent d'étudier les réactions biochimiques qui se réalisent entre un substrat et son produit.

Une fois que les cartes représentants les différentes interactions cellulaires sont rassemblées en un réseau, il est possible de l'exploiter en analysant la dynamique de ce réseau. Pour cela deux formalismes sont applicables. Tous les deux utilisent les connaissances sous forme d'équation et de règles pour calculer les valeurs de noeuds du réseaux par rapport à celles des autres noeuds.

Le premier est basé sur des équations différentielles (*Ordinary Differential Equation* ou ODE). La construction de ces équations requiert de nombreuses valeurs stoechiométriques pour caractériser l'état du noeud comme les constantes de diffusion. Ces éléments sont difficiles à obtenir puisqu'ils nécessitent des conditions *in vitro* très proches de la réalité biologique [20],

4. https ://reactome.org/
5. https ://www.genome.jp/kegg/
6. https ://www.pathwaycommons.org/

[21].

Le deuxième formalisme va simplifier ces règles. Les valeurs de chaque noeud ne sont plus des valeurs continues mais booléennes [22]. Elles traduisent la notion d'activation/inactivation dans les réseaux de signalisation par exemple, d'expression ou non dans les GRN. La dynamique de ces réseaux booléens est régie par des formules logiques pour chacun des noeuds. Elles peuvent être extraites de base de données ou inférées via la littérature. Une formule d'un noeud traduit l'interaction des noeuds précédents via des portes logiques (AND, OR, NOT). L'utilisation de valeurs booléennes pour caractériser les noeuds peut être réductrice par rapport à la réalité biologique. On peut donc étendre ce concept en ajoutant des seuils, en remplaçant les valeurs 0 et 1 par des valeurs multi-valuées [23], [24].

Dans les deux cas, la connaissance de cette dynamique entre les éléments biologiques permet d'identifier de manière automatique les dépendances de régulation entre les noeuds. Ainsi, on peut apprendre quelles sont les entités biologiques nécessaires pour l'activation d'une voie particulière par exemple. Lorsque le système biologique arrive à son état d'équilibre (état stable) et qu'il s'agit d'un phénotype d'intérêt, on peut extraire les gènes ou protéines qui lui sont constamment associés [25]. Ces éléments ne seraient peut être pas identifiés en expérimentation biologique si leur expression n'était pas suffisamment importante.

Dans les GRN, ces états stables peuvent être une représentation des valeurs obtenues dans des expériences de transcriptomique. L'analyse de la dynamique pourrait donner les valeurs de certains noeuds (exprimés ou non) qui sont nécessaires pour atteindre ces états d'équilibre. Les éléments de cause d'un phénotype seraient ainsi extraits des éléments de conséquence liés à l'activation de certaines voies.

La recherche de causalité grâce à cette approche est aussi utilisée dans une partie de la thèse.

Dans le cadre de cette thèse, ces deux approches sont appliquées sur une maladie auto-immune systémique particulière. Il s'agit du lupus systémique erythémateux. Il fait partie des MAI recensées dans le consortium PRECISESADS.

III Le lupus systémique erythémateux, une pathologie complexe et hétérogène

Le lupus systémique erythémateux (*Systemic Lupus Erythematosus* ou SLE) fait partie de cette catégorie de maladie complexes influencées par plusieurs facteurs génétiques et environnementaux. D'après la Filière de Santé des Maladies Auto-Immunes et AutoInflammatoires Rares (FAI2R [7]), le lupus est considéré comme une maladie rare car sa prévalence est inférieure à 50 cas pour 100 000 habitants (environ 40/100 000). De plus, le lupus a tendance à toucher plus

7. https ://www.fai2r.org/

les femmes que les hommes avec un ratio de 9 pour 1. Les femmes les plus atteintes ont entre 15 et 45 ans, période où les hormones ont le plus de fluctuation avec le cycle menstruel. Ce qui laisse supposer un rôle des hormones dans le déclenchement de la pathologie.

III.i Symptômes et diagnostic

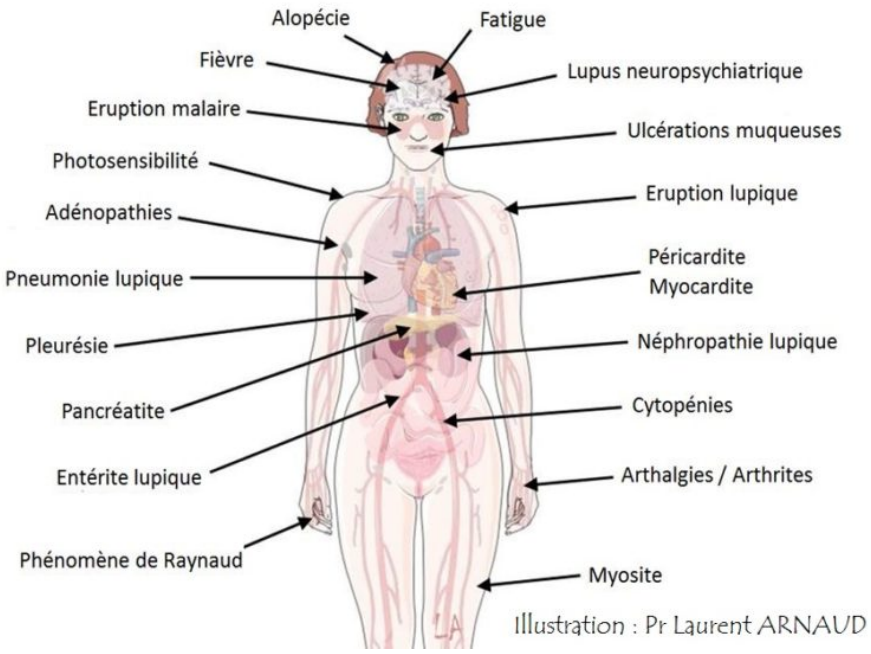


FIGURE 4 – Schéma représentant les différentes manifestations associées au SLE (issue de [a])

a. https ://maladie-autoimmune.fr/lupus-systemique-maladie-auto-immune/

La complexité de la maladie est un facteur important dans l'hétérogénéité de ses symptômes. De par sa nature systémique, plusieurs organes peuvent être touchés, mais ce ne sont pas forcément les mêmes pour chaque patient. La figure 4 issue du site du CRESO, monte les différentes atteintes provoquées par la maladie. De plus, la pathologie n'est pas caractérisée par une évolution chronique, mais par des successions de poussées puis de rémissions. Ces crises peuvent être plus ou moins sévères. Elles se traduisent par l'apparition de simples rashs sur la peau jusqu'à une forme plus sévère [26]. Le pronostic vital est notamment engagé lors d'une atteinte viscérale comme une insuffisance rénale (d'après le Protocole National de Diagnostic et de Soins de la Haute Autorité de Santé [8]).

Afin de diagnostiquer cette maladie, plusieurs aspects sont mesurés : la présence d'auto-anticorps et la sévérité de l'atteinte des organes.

8. https ://www.has-sante.fr/

Afin de mesurer l'atteinte des organes, deux scores sont utilisés en clinique. Le premier est appelé score SLEDAI (SLE Disease Activity Index). Il mesure l'activité de la pathologie et est basé sur plusieurs observations issues d'un consensus entre cliniciens [27]. Ce score mesure 24 aspects biologiques et cliniques pour chaque patient et attribue une valeur à chacun de ces aspects. Le score fluctue ainsi entre 0 et 105 et a, par la suite, été dérivé en 3 classes. Lorsque le score est entre 0 et 6, l'activité pathologique du patient est dans la catégorie minime (*mild*), entre 6 et 9 dans la catégorie modérée (*Moderate*) et plus de 9 dans la catégorie sévère (*Severe*).

Le second score est le score BILAG (*British Isles Lupus Assessment Group*) utilisé aussi pour mesurer l'activité de la maladie [28]. Sa mesure est réalisée sur 84 aspects répartis sur 9 domaines ou organes. Pour chacun de ces aspects, un score entre A et E est attribué selon la sévérité des symptômes. De la même manière que pour le score SLEDAI, 3 grandes classes ont été définies pour résumer le stade de la pathologie : maladie active ou s'aggravant, maladie s'améliorant, maladie à persistance minimale ou sans activité.

Ces deux scores reposent sur l'observation des patients par les cliniciens. Même avec leur expertise, il subsiste une part de subjectivité dans l'attribution de ces valeurs. De plus, l'hétérogénéité des symptômes combinés à l'historique des patients s'explique par la complexité de cette maladie.

Comme pour toutes les MAI, il n'existe pas à l'heure actuelle de cause connue. Cependant, les mécanismes moléculaires impliqués sont fortement liés aux cellules lymphocytaires.

III.ii Rôle des lymphocytes B et T et de l'interféron

Tout comme les MAI de manière générale, l'étiologie du SLE est encore peu comprise. Cependant, le dysfonctionnement du système immunitaire est associé avec le développement de la maladie.

Wahren-Herlenius et Dörner ont résumé les principales voies de dérégulation dans le SLE [29] et la figure 5 représente les différents mécanismes impliqués dans la pathologie [30].

Les lymphocytes B et T ont un rôle prépondérant dans l'immunité adaptative. D'une part, les lymphocytes T peuvent se différencier en un sous-ensemble d'effecteurs : CD4+ helper (Th), CD4+ regulateurs (Treg), CD8+ cytotoxique. Dans le contexte du SLE, les Th et les Treg ont été montrés comme ayant une implication particulière dans la pathologie. Les patients SLE présentent une augmentation du nombre de cellules Th17 qui jouent un rôle pro-inflammatoire et induit des dommages aux tissus. En revanche, les Treg ont un rôle dans le maintien de la tolérance des lymphocytes T et B. Les patients SLE ont donc tendance à avoir une diminution de l'abondance de ces cellules. D'autre part, l'auto-réactivité des lymphocytes B va entraîner la production d'auto-anticorps, marqueurs cliniques de la pathologie.

Ces dérégulations peuvent être provoquées par des polymorphismes dans les cellules lymphocytaires mais aussi par l'immunité innée. Des cellules immunitaires comme les cellules den-

dritiques (DC), les plasmocytes (Pla) et les macrophages vont médier l'activation de la réponse immunitaire par la production de cytokine comme les interleukines (IL) et surtout par celle des interférons (IFN).

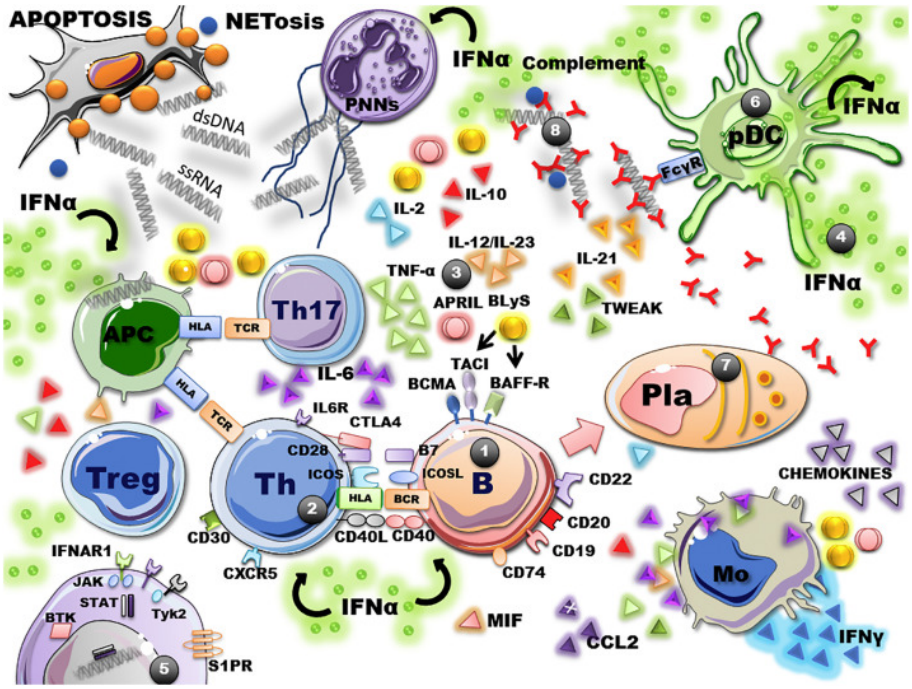


FIGURE 5 – Représentation des mécanismes impliqués dans le SLE ainsi que des cibles thérapeutiques potentielles (issue de [30])

Le SLE fait partie des premières pathologies auto-immunes pour lesquelles l'interféron (IFN) a été identifié comme facteur important. Il a d'ailleurs souvent été corrélé avec l'activité clinique de la pathologie. C'est en 2003 qu'un motif d'expression de gènes a été identifié, désormais défini comme étant la signature IFN [31]. Cette signature est basée sur des analyses à haut-débit d'expression de gènes qui est mesurée en comparant une population de patients SLE contre une population d'individus sains. Elle est composée de 29 gènes, majoritairement sur-exprimés chez les patients SLE. Beaucoup de ces gènes ont une expression induite par la voie de l'IFN comme IFIT3, IFI44 et IFI44L, ce qui favorise l'hypothèse d'un rôle important de l'IFN dans le développement de la pathologie.

Depuis, de nombreuses études ont confirmé la présence de cette signature dans les cohortes de patients SLE. Les analyses d'enrichissement montrent aussi la présence de gènes associés à la voie de signalisation de l'IFN de type 1 et de l'IFN α ainsi qu'aux réponses de l'immunité innée [32], [33].

Á partir de ce postulat, plusieurs traitements ont été développés ciblant les voies de l'IFN et les cellules lymphocytaires.

III.iii Traitement thérapeutique

Á l'heure actuelle, il n'existe pas de traitement permettant de guérir la maladie. Mais plusieurs approches ont été développées afin de soigner les symptômes, comme la corticothérapie, l'utilisation d'hydroxychloroquine ou dans des cas extrêmes, des immunosuppresseurs [34].

L'objectif de nombreuses études est de développer plus de thérapies permettant d'enrayer les mécanismes moléculaires de la pathologie. Ainsi, 3 grands groupes de thérapie ont été résumés par Touma et Gladman selon la cible : l'activation des lymphocytes B, celle des lymphocytes T et la voie de l'IFN [35].

Le seul médicament ayant eu une autorisation de mise sur le marché (AMM) est le Belimumab (GSK) [36]. Cet anticorps monoclonal agit sur le facteur d'activation des lymphocytes B (BAFF) et réduit le nombre de lymphocytes B. Les autres thérapies ciblées sont actuellement en essais cliniques.

C'est le cas de :

— Atacicept (Merck) : anticorps monoclonal ciblant les molécules BAFF et APRIL
— Abatacept (Orencia) : inhibiteur de l'activation des lymphocytes T
— Anifrolumab (AstraZeneca) : antagoniste du récepteur de l'IFN-α

Plusieurs développements d'autres médicaments n'ont malheureusement pas pu aboutir au cours des essais cliniques. Comme le souligne Touma et Gladman, le problème réside dans l'hétérogénéité des patients face à cette maladie combinée à la petite taille d'échantillons souvent retrouvée dans les études cliniques.

Les contributions de ma thèse sont réparties en 4 chapitres. Chaque chapitre traite d'une problématique particulière. Mon choix a donc été de positionner chaque problématique par rapport à son état de l'art plutôt que de manière générale dans l'introduction.

IV Résumé des contributions de la thèse

Le premier chapitre présente l'étude clinique du SLE mise à disposition par SANOFI. **La première contribution** a été d'identifier le type d'information à extraire de ces données. Tout d'abord, le concept de signature pathologique a été identifié à partir d'une recherche bibliographique et trois catégories de signatures (diagnostic, pronostic et prédictive) sont ressorties. Ces signatures sont généralement basées sur une analyse statistique avec des données d'expression de gènes. Le contexte de l'étude nécessite d'utiliser le principe de signature de diagnostic qui est normalement utilisée pour définir les éléments propres à une population malade. Des analyses différentielles de l'expression sont réalisées en utilisant plusieurs caractéristiques pathologiques (malade ou non, score SLEDAI).

Les résultats ont montré les limites de ce type d'analyse. La première étant que la compa-

raison de populations n'était pas appropriée car les patients SLE sont très hétérogènes et que le nombre d'individus est très inférieur au nombre de variables explicatives. La deuxième limite est le fait que toutes les données omiques ne peuvent pas être prises en compte. Les données de génotypage étant des données discrètes, elles sont difficilement utilisables en statistique. La troisième limite est le fait que les résultats issus de l'analyse d'expression de gènes représentent un mélange d'éléments impliqués dans la cause de la maladie, ses conséquences ainsi que du bruit lié à la population.

L'objectif de la thèse s'est donc orienté vers la définition d'une méthode intégrant les aspects massifs et hétérogène des données omiques, l'hétérogénéité de la population et la recherche de causalité pour un phénotype d'intérêt.

La problématique du chapitre 2 est celle de l'intégration des données massives et hétérogènes que présente l'étude clinique. L'état de l'art montre qu'il n'existe pas à l'heure actuelle de schéma d'unification et de connexion pour ces données. **Ma deuxième contribution** a donc été de définir une structure pour les données permettant de relier les valeurs cliniques des individus, leur mesure d'expression et leur génotypage avec les caractéristiques des gènes. Pour cela, la représentation en RDF des données a été utilisée. Un schéma a été défini et est généralisable à toutes les études cliniques ayant les mêmes types de données omiques. Le concept de modèle transomique est défini dans ce cas, puisqu'il permet de connecter plusieurs niveaux biologiques.

L'alimentation de ce modèle nécessite de prendre en compte l'hétérogénéité de la population. L'état de l'art montre que les analyses des données multi-omiques sont principalement des analyses statistiques ou de machine learning reposant sur l'utilisation de comparaisons entre les populations. L'originalité de ma contribution a été d'utiliser une approche centrée sur le patient. Dans les méthodes utilisant ce type d'approche, la personnalisation des données se base sur l'utilisation de réseaux biologiques, extraits de bases de données ou sur des réseaux de co-expression qui ne reflètent pas les liens de régulation biologiques. **Ma troisième contribution** est la proposition d'une méthode d'alimentation du modèle transomique permettant de transformer les données brutes de l'étude clinique en triplestore pour lesquels les valeurs d'expression de gènes et de présence de SNPs sont spécifiques à chaque patient.

Ces deux contributions seront appliquées sur une étude clinique présentant les mêmes caractéristiques que celle de SANOFI. Plusieurs requêtes SPARQL simples seront construites afin de valider l'intégration des données.

Le chapitre 3 présente un nouveau type de liens de causalité potentiel entre la génomique et la transcriptomique à l'échelle d'un patient. **Ma contribution dans ce chapitre** est d'avoir défini formellement ce lien et de l'avoir traduit en une requête SPARQL complexe qui interroge

le modèle transomique défini dans le chapitre précédent. Un eICTL est un couple SNP-gène pour lequel la présence d'un SNP influence la variation d'expression de son gène chez au moins deux patients. Cette définition a été traduite en langage SPARQL et appliquée au modèle transomique. Les eICTLs identifiés ont ensuite été évalués en comparant leur apport avec les données de génomique.

La recherche de causalité ne se limite pas à l'utilisation des données omiques et des liens généraux entre eux. Il est aussi important de prendre en compte les liens entre les entités.

Ma contribution dans le chapitre 4 est d'utiliser la dynamique d'un système booléen biologique (réseaux de régulation de gènes) afin d'enrichir des signatures biologiques déjà établies. Pour cela, l'outil GINsim a été utilisé pour calculer l'ensemble des états stables de réseaux booléens représentant la différenciation des lymphocytes T. Les différents phénotypes des lymphocytes ont été préalablement associés à l'expression, l'activation d'un noeud spécifique du réseau. Cet élément est considéré comme étant le régulateur majeur de ce phénotype et constitue donc sa signature. En utilisant la méthode d'analyse par concept formel, l'ensemble des états stables de ces réseaux a pu être classé en différents phénotypes selon les valeurs de leur noeuds. Des motifs d'activation de noeuds ont pu être identifiés dans ces différents groupes entraînant l'enrichissement des signatures définies au départ. Les noeuds supplémentaires peuvent correspondre à des éléments causaux dont la présence est nécessaire à l'expression du régulateur majeur ou favorise la différenciation cellulaire.

Cette contribution a fait l'objet d'une publication dans le *Journal of Theoretical Biology* en 2019.

Chapitre 1

Signature populationnelle dans le cadre d'une maladie complexe

La recherche de marqueurs biologiques pour une pathologie d'intérêt passe par l'analyse de patients dans un contexte d'études cliniques. Ce type d'analyse dans un contexte industriel permet l'identification de cibles thérapeutiques potentielles. La collaboration avec l'entreprise pharmaceutique SANOFI a permis d'avoir à disposition une étude clinique longitudinale pour la thèse. Cette étude porte sur le lupus systémique érythémateux.

Ce premier chapitre présente les premières analyses des données réalisées pendant la thèse. La première section du chapitre est une présentation de l'étude clinique sur le SLE de SANOFI. La deuxième section présente les différentes définitions du concept de signature dans un contexte biomédical qui sont extraites d'une analyse bibliographiques. En troisième section, des gènes différentiellement exprimés de l'étude clinique sont mesurés par une approche statistique. Les plus significatifs sont utilisés comme potentiels motifs de classification de la population SLE. Cependant, les résultats montrent que cette signature n'est pas suffisante pour stratifier les patients SLE. La dernière section conclut sur les deux limites qui ont été soulevées à la suite des résultats. D'une part, l'hétérogénéité de la population de patients SLE et d'autre part, le mélange de cause, de conséquence et de bruit dans les mesures issues de données d'expression. Enfin, l'objectif de la thèse et l'approche proposée pour pallier ces limites sont présentés.

1.1 Présentation de l'étude clinique sur le lupus systémique érythémateux (SLE) (données SANOFI)

Mon travail de thèse repose sur une étude clinique longitudinale menée par SANOFI Genzyme. 191 individus ont été suivis sur une année : 171 sont des patients SLE (population d'intérêt) et 20 sont des individus sains (population contrôle). La première visite correspond au début de l'étude et sert de temps 1 des patients SLE (*baseline*) pour le reste de l'étude. Un rendez-vous est pris tous les 3 mois par la suite, afin de réaliser un prélèvement sanguin pour les analyses biologiques ainsi qu'un suivi des symptômes. Cela représente un maximum de 5 visites (baseline + 4 visites) par patient SLE. Cependant, il s'agit d'une étude clinique observationnelle. Rien

n'oblige les patients à se présenter à chaque visite afin d'obtenir un traitement. Le design expérimental est donc incomplet. Certains patients sont venus à chaque visite, d'autres à quelques unes mais de manière irrégulière. Quand aux individus sains, ils ont 1 à 2 visites sur l'ensemble de l'étude.

Plusieurs analyses biologiques ont été réalisées et sont issues de prélèvements sanguins. L'expression des gènes a été mesurée par RNAseq pour tous les individus, à chaque visite. Le génotypage a été réalisé une fois *via* une puce à ADN pour uniquement 168 individus. Il y a donc 23 patients pour lesquels l'information de génotypage n'est pas connue. Des analyses de cytométrie en flux ont été effectuées pour identifier les populations cellulaires immunitaires dans les échantillons ainsi que la recherche de biomarqueurs. Enfin, la quantification des anti-corps, considérés comme biomarqueurs de la pathologie et contenus dans le sérum, a été calculée par puce à auto-anticorps. L'ensemble de ces analyses a été réalisé sur du sang complet.

Les données cliniques enregistrées permettent de stratifier les patients selon leur état pathologique (sain ou malade). Les scores SLEDAI et BILAG sont aussi renseignés, ce qui permet de mesurer le degré de la maladie. Comme présenté dans l'introduction, le SLE est une maladie qui se présente sous forme de crise. Ainsi, les temps de visite ne correspondent pas forcément à une période de crise. La figure 1.1 nous permet de voir que le score SLEDAI de chaque patient varie donc selon leur visite.

L'évolution d'une pathologie peut se faire de manière progressive avec des stades plus ou moins agressifs qui se succèdent, comme c'est le cas des cancers solides. En revanche, comme expliqué dans l'introduction, le SLE est une maladie qui apparaît sous forme de crise. Le diagramme d'états des échantillons de l'étude clinique (Figure 1.2) montre les probabilités de changement de catégories entre deux visites pour les patients SLE. Pour chaque catégorie, la probabilité la plus grande est celle de rester dans la même catégorie : 0.76 pour la catégorie "mild", 0.44 pour "moderate" et 0.46 pour "severe". Cette représentation montre aussi que toutes les transitions sont possibles. Un patient peut passer d'une catégorie "severe" à une catégorie "mild" (probabilité de 0.2) et inversement (probabilité de 0.05). Pour la catégorie "moderate", les patients tendent à avoir des crises moins agressives à la visite suivante (probabilité de 0.41 vers la catégorie "mild" contre 0.14 vers la catégorie "severe"). Grâce à ce graphe, on peut voir qu'il n'y a pas de relation entre les états cliniques des patients au cours des visites. Les crises ne semblent pas organisées dans un enchaînement bien particulier.

L'objectif de cette étude clinique est de suivre, sur un an, certains biomarqueurs connus pour être associés à la maladie. Dans le cadre de ma thèse, le but a été de pouvoir identifier un motif d'éléments biologiques et cliniques pouvant expliquer la maladie. A la différence des approches classiques qui sont décrites dans la section suivante, cette identification passe par une recherche dans les signaux faibles des données, ceux spécifiques de sous-ensemble de patients et non pas dans les signaux retrouvés dans toute une population d'intérêt. Les données contenant le plus

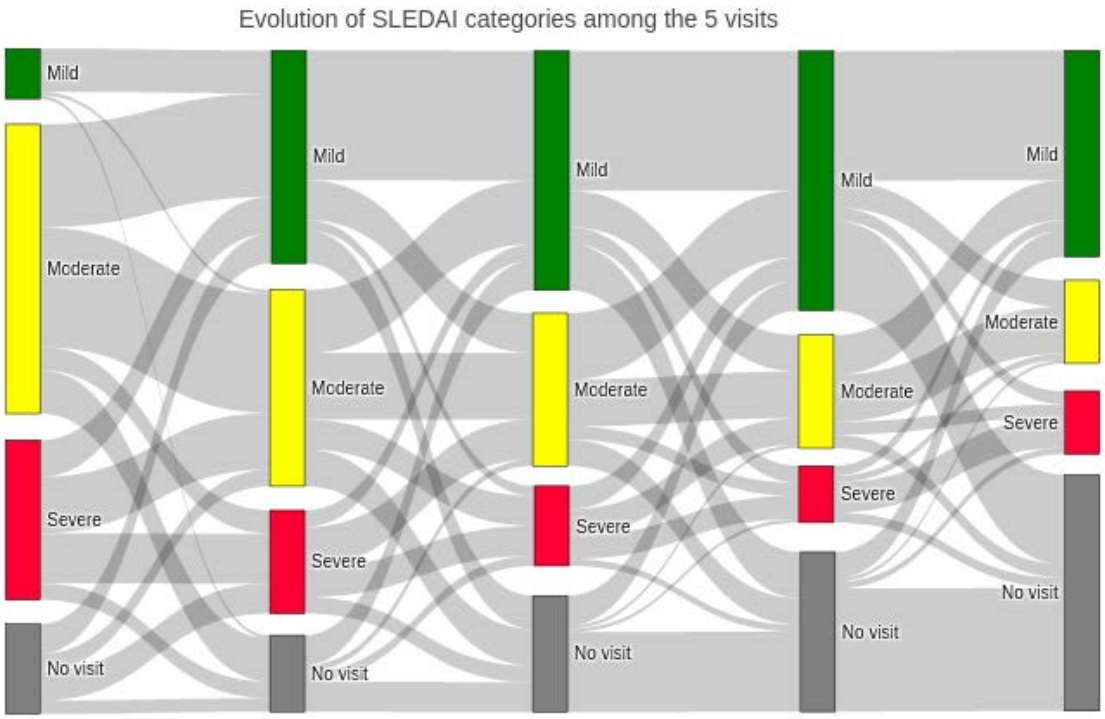


FIGURE 1.1 – **Diagramme de Sankey représentant l'évolution du score SLEDAI sur les 5 visites**. Chaque colonne représente une visite (V1, V2, V3, V4, V5). Pour chaque visite, les patients sont classés en 4 catégories selon leur score SLEDAI (Mild, Moderate, Severe ou pas de visite). La taille de chaque catégorie par visite représente la proportion de patients par catégorie. Les arcs entre les visites correspondent à la proportion de patients qui sont passés d'une catégorie X à une visite t à une catégorie Y à la visite $t+1$.

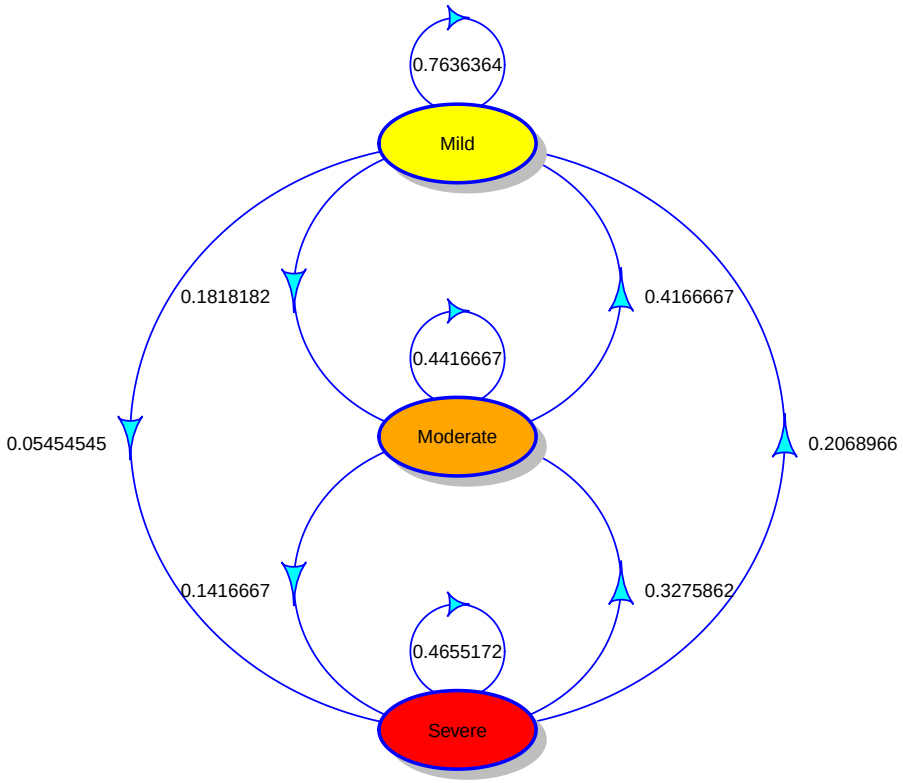


FIGURE 1.2 – **Transition d'état entre les catégories SLEDAI**. Chaque ellipse correspond à une catégorie du score SLEDAI. Les arcs représentent les transitions entre les catégories SLEDAI et les chiffres sont les probabilités qu'un patient change de catégorie ou non à la visite *t+1* sachant leur catégorie à la visite *t*.

d'éléments, génomique, transcriptomique et clinique ont été récupérées.

1.2 Différentes signatures pour différents contextes pathologiques

Dans l'étude de maladies, de nombreuses études portent sur l'identification de signature. Elles s'accordent toutes pour décrire une signature comme étant un ensemble de marqueurs (gène, protéine, métabolite) qui contient une information utile [37]-[41]. L'objectif est d'obtenir une liste d'éléments biologiques associés à un phénotype d'intérêt. Elle peut être associée à un paramètre clinique particulier ou permettre de prédire un phénotype clinique d'intérêt.

Une signature s'obtient par comparaison entre deux populations caractérisée par des aspects phénotypiques propres où les éléments qui composent la signature sont spécifiques à l'une des populations. L'obtention d'une signature est donc conditionnée à la stratification préalable des individus (Tableau 1.1).

L'identification d'une signature peut se résumer en quelques étapes [39], [42]. Après avoir défini le contexte clinique et scientifique (étape 1) et pré-processé les données (étape 2), il est nécessaire de réaliser une sélection des éléments importants et de construire le modèle statistique associé (étape 3). Le type des composants de cette signature est dépendant des données analysées. Il peut s'agir d'une expression génique, de méthylation, de CNV, d'abondance de protéines ou de métabolites. Différents modèles statistiques peuvent être appliqués selon le type de données ainsi que le type de comparaison entre les populations. Selon les méthodes, il est possible d'intégrer des connaissances extraites des bases de données telles que KEGG ou Reactome. Une fois le motif moléculaire identifié, il faut évaluer cette signature. Il s'agit de vérifier que le motif n'est représenté que dans une seule des deux populations. Le clustering, les analyses d'enrichissement et les forêts aléatoires (*random forest*) sont, par exemple, des approches qui peuvent être utilisées pour l'étape d'évaluation. Enfin, il est nécessaire de valider le motif sur une cohorte qui ne soit pas celle où la signature a été calculée.
L'interprétation biologique de cette signature dépend aussi du contexte biologique associé à ces données. On peut identifier 3 grandes catégories de signature [41], [43].

Type de signature	Type de population	Référence
Diagnostic	Sains *vs* Malades	[44]
Pronostic	Malades à un stade non-agressif *vs* agressif	[45], [46]
Prédictive	Malades traités *vs* non-traités	[47], [48]

TABLE 1.1 – **Tableau récapitulatif des différentes signatures pathologiques : diagnostic, pronostic et prédictive.** Chaque signature est associée au type de population utilisé dans la comparaison ainsi qu'aux références dans le contexte du SLE

Une signature de *diagnostic* met en évidence des caractéristiques moléculaires qui sont spécifiques à une pathologie [49] et peut être utilisée par la suite pour identifier les personnes

malades. Certaines molécules identifiées peuvent par la suite servir de nouvelles cibles thérapeutiques. Dans les cas de cancer solide par exemple, il est possible de comparer du tissu sain et du tissu tumoral d'un même patient pour identifier des marqueurs cancéreux. Cette comparaison prend en compte l'hétérogénéité intra-individuelle des patients. En revanche, le SLE étant une maladie diffuse qui touche l'ensemble des organes, il n'est pas possible de distinguer un tissu sain d'un tissu malade pour un même patient. En 2003, une première signature de SLE, contenant des gènes inductibles par l'interféron, a été identifiée en comparant les expressions de gènes, mesurées par micro-array dans des cellules mononucléées sanguines périphériques (ou *PBMCs*), d'une population de patients SLE à ceux d'une population saine [44].

La signature de *pronostic* se calcule sur des populations malades uniquement [40], [50], [51]. La signature met en avant les éléments moléculaires associés à une meilleure survie des patients ou à une modification de l'activité de la maladie. Elle sera calculée entre les différentes stades d'un cancer par exemple. Dans le cas du SLE, l'augmentation de l'acétylation des glycoprotéines [45] ou de la concentration en granzyme B, des protéases à sérine, [46] ont été récemment identifiées comme étant des signatures de pronostic, puisqu'elles augmentent chez les patients SLE dont le score SLEDAI est plus important que chez les autres patients.

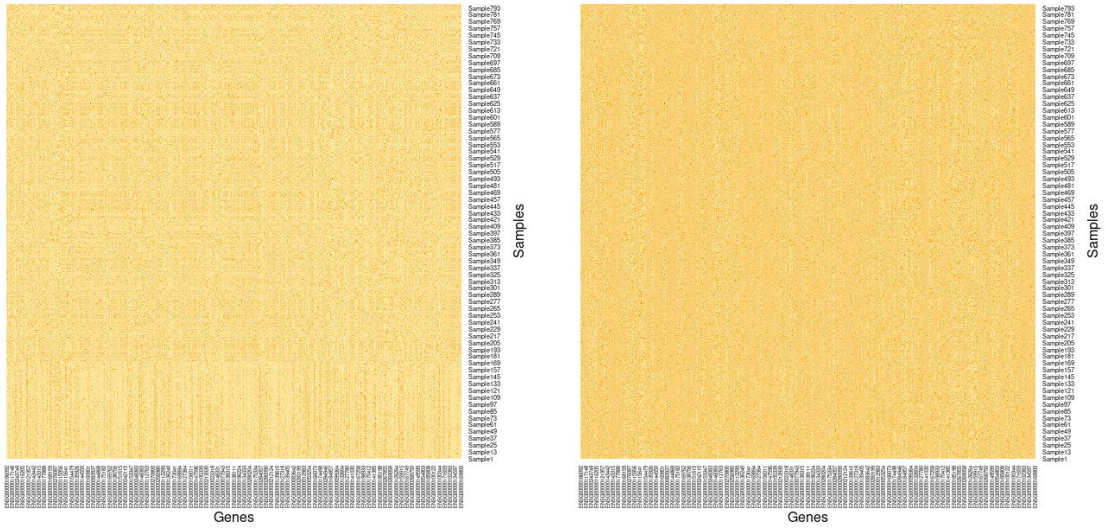
Elle diffère de la signature *prédictive* qui permet de définir un traitement optimal selon les caractéristiques biologiques du patient [51], [52]. En effet, certaines sous-populations de patients peuvent être plus sensibles à un traitement ou ne peuvent pas, à l'inverse, le supporter. C'est le cas de certaines mutations qui empêchent la métabolisation du principe actif et entraînent des effets secondaires importants [47]. On parle alors de signature pharmacogénomique [53]. Bertolo *et al.* ont comparé, après 6 mois d'une thérapie d'induction, les patients SLE qui répondaient bien au traitement, avec ceux dont la condition ne présentait aucune évolution [48]. Ils ont montré qu'une forte concentration des lymphocytes T CD4+ effecteurs à mémoire dans les urines permettait de prédire une réponse insuffisante à cette thérapie d'induction.

Beaucoup de ces différentes signatures se basent sur des données d'expression comme les techniques de microarray et de RNA-seq. Comme présenté dans la section précédente, l'étude clinique de SANOFI implique deux types de population, des individus sains et des patients SLE. La définition de signature de diagnostic se rapproche plus du contexte de la thèse.

1.3 Recherche de signature de diagnostic dans l'étude clinique

La première étape de la thèse a été de réaliser une analyse comparative des données d'expression entre la population contrôle, les individus sains, et la population d'intérêt, les patients SLE. Les signatures de diagnostic basées sur des données transcriptomiques sont composées d'un ensemble de gènes différentiellement exprimés (DEG). Pour cela, l'ensemble des échantillons a été utilisé, représentant 19 883 gènes mesurés dans 800 échantillons.

Une première visualisation des données d'expression a montré un effet batch entre les échantillons mesurés au premier jour de l'étude et le reste des mesures (Figure 1.3). Cette différenciation peut être due à une différence technique entre les échantillons prélevés ou analysés au début de l'étude et ceux du reste de l'étude. Cet effet a été supprimé par la fonction *removeBatchEffet* du package R *limma* pour la visualisation des heatmaps et sera prise en compte comme co-variable dans le reste de l'analyse des DEGs.



(a) Heatmap des données d'expression des patients SLE avec l'effet batch

(b) Heatmap des données d'expression des patients SLE sans l'effet batch

FIGURE 1.3 – **Comparaison des données d'expression avec et sans effet batch.** Pour chaque heatmap, les lignes correspondent aux échantillons et les colonnes aux gènes mesurés.

Les calculs de DEG ont été réalisés *via* le package *edgeR*. A partir des comptages normalisés, l'analyse identifie les DEG (dont la p-value corrigée est inférieure à 5% et dont la valeur absolue du fold-change (FC) est supérieure à 1.5). Cela signifie que les gènes identifiés sont 1.5 fois plus (ou moins) exprimés chez les patients SLE que chez les individus sains et ce, de manière significative. Un deuxième filtre plus stringent a été appliqué en ne gardant cette fois que les gènes dont le ratio d'expression est 2 fois plus (ou moins) important chez les patients SLE.

L'ensemble des résultats de calcul des gènes différentiellement exprimés est présenté dans les sous-sections suivantes.

1.3.1 Comparaison entre la stratification des patients selon les catégories SLEDAI et la population saine

Les scores cliniques sont utilisés pour stratifier plus finement la population malade. Les 3 catégories du score SLEDAI sont donc utilisées de manière séquentielle, pour identifier des gènes qui pourraient être spécifiques à chaque catégorie (Figure 1.2). La comparaison de la première

Comparaison	FDR <5%, ||FC|| > 1.5	FDR <5%, ||FC|| > 2
SLEDAI Mild *vs* sains	795	400
SLEDAI Moderate *vs* sains	429	298
SLEDAI Severe *vs* sains	525	259

TABLE 1.2 – **Analyse des gènes différentiellement exprimés selon score SLEDAI des patients de l'étude clinique SANOFI par *EdgeR* (package R).** Une comparaison a été réalisée avec une stratification fine des patients en comparant les échantillons de chaque catégorie SLEDAI par rapport aux contôles. Pour chaque analyse, un filtre pour la p-value corrigée est appliquée (inférieure à 0.05) ainsi que pour la différence d'expression (fold-change) : soit 1.5 fois, soit 2 fois différent.

catégorie, "mild" avec le groupe contrôle identifie 795 DEG. Celle avec la catégorie "moderate" contre le groupe contrôle en identifie 429. Avec la catégorie "severe" contre le groupe contrôle, 525 DEGs. Avec le deuxième filtre sur le FC, on obtient 400 pour la catégorie "mild", 298 pour la catégorie "moderate" et enfin 525 pour la catégorie "severe".

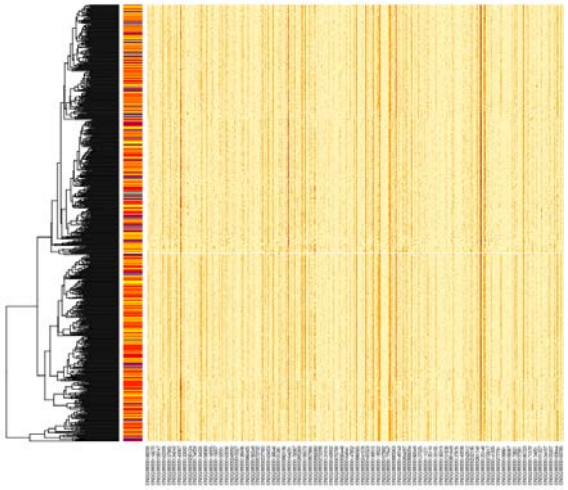
Ces résultats montrent que de nombreux éléments sont identifiés avec les deux filtres puisqu'un minimum de 259 DEGs sont retrouvés. La catégorie "mild" est celle ayant le plus d'éléments comparée aux deux autres catégories. Ce grand nombre d'éléments peut s'expliquer par le petit nombre d'échantillons SLE (n) dans les catégories par rapport au grand nombre de variables explicatives (p), diminuant la puissance statistique de cette approche.

Une fois que les gènes dont l'expression est la plus marquée chez les patients ont été sélectionnés, il est nécessaire d'évaluer le potentiel de classification de cette signature. Pour cela, une matrice est créée en prenant l'ensemble des individus et l'union de l'expression des DEGs identifiés entre les sains contre chaque catégorie SLEDAI. Cette évaluation est réalisée par clustering hiérarchique des patients selon ces différents éléments.

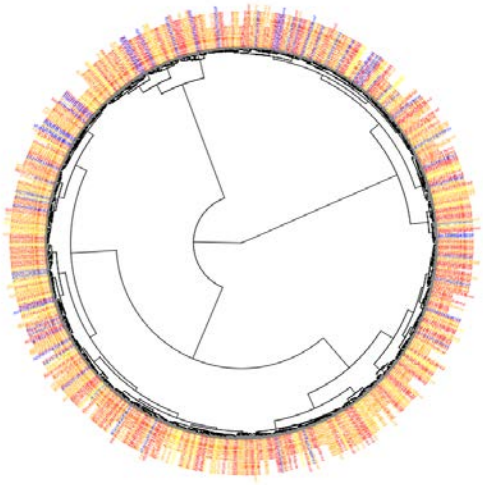
L'évaluation de la signature censée caractériser les différentes catégories SLEDAI est présentée en figures 1.4a et 1.4b. Les lignes de la heatmap représentent les échantillons avec leur score SLEDAI selon différentes couleurs : bleu pour échantillon sain, jaune pour la catégorie "mild", orange pour "moderate" et rouge pour "severe" (figure 1.4a). La réorganisation des échantillons selon leur similarité montre que les gènes identifiés comme différentiellement exprimés pour chaque catégorie ne sont pas suffisamment informatifs. En effet, le regroupement des patients en fonction de la proximité de leur signature ne permet pas de retrouver la classification d'origine, comme le montre le dendrogramme en figure 1.4b.

1.3.2 Comparaison entre la population malade et la population saine

Le score SLEDAI étant un score observationnel, il peut exister de l'hétérogénéité entre les patients selon le clinicien. Une seconde analyse statistique qui compare les malades avec les sains identifie un ensemble de 440 DEG avec un filtre de FC à 1.5 et 199 pour un filtre de



(a) Heatmap des échantillons en utilisant la matrice des DEG des signatures des différentes catégories SLEDAI



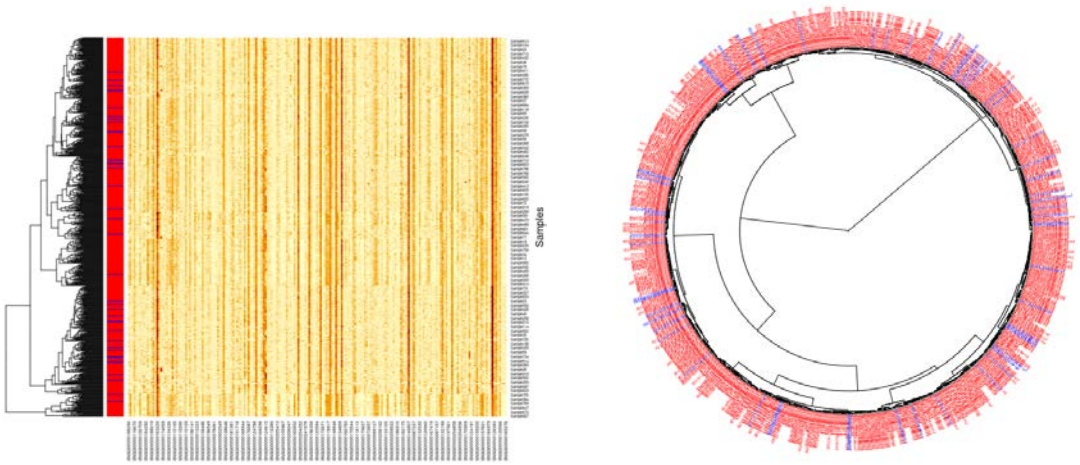
(b) Dendrogramme des échantillons en utilisant la matrice des DEG des signatures des différentes catégories SLEDAI

FIGURE 1.4 – **Evaluation des DEGs identifiés selon les catégories SLEDAI et les contrôles.** Pour chaque heatmap, les lignes correspondent aux échantillons et les colonnes aux DEGs identifiés. Pour les heatmaps et les dendrogrammes, les échantillons sont organisés selon leur similarité. (a) (b) Graphes représentent l'évaluation de la combinaison des DEGs identifiés dans les comparaisons des malades avec les différents scores SLEDAI par rapport la population contrôle.

Comparaison	FDR <5%, $\|\|FC\|\| > 1.5$	FDR <5%, $\|\|FC\|\| > 2$
Malade SLE *vs* sains	440	199

TABLE 1.3 – **Analyse des gènes différentiellement exprimés entre les sains et les malades de l'étude clinique SANOFI par *EdgeR* (package R).** Une comparaison plus large a été faite en comparant les malades par rapport aux sains. Pour chaque analyse, un filtre pour la p-value corrigée est appliquée (inférieure à 0.05) ainsi que pour la différence d'expression (fold-change) : soit 1.5 fois, soit 2 fois différent.

FC à 2 (Figure 1.3). Dans ce cas, le nombre de patients SLE est plus important par rapport aux analyses précédentes ce qui entraîne une augmentation des contraintes appliquées sur les variables explicatives.



(a) Heatmap des échantillons en utilisant la matrice des DEG de la signature SLE

(b) Dendrogramme des échantillons en utilisant la matrice des DEG de la signature SLE

FIGURE 1.5 – **Evaluation des DEGs identifiés chez tous les malades et les contrôles.** Pour chaque heatmap, les lignes correspondent aux échantillons et les colonnes aux DEGs identifiés. Pour les heatmaps et les dendrogrammes, les échantillons sont organisés selon leur similarité. (a) (b) Graphes représentent l'évaluation de la combinaison des DEGs identifiés dans les comparaisons des malades avec les différents scores SLEDAI par rapport la population contrôle. (c) (d) Graphes représentent l'évaluation des DEGs identifiés dans la comparaison des malades avec la population contrôle.

Une seconde évaluation a donc été réalisée à un niveau plus large, en utilisant les gènes identifiés dans la comparaison des échantillons sains avec les échantillons SLE (figures 1.5a et 1.5b). Tout comme l'évaluation précédente, les lignes de la heatmap représentent les échantillons qui ont été réordonnés selon leur similarité. Les échantillons sains sont représentés en bleu alors que les échantillons SLE sont en rouge (figure 1.5a). Sur la heatmap ainsi que sur le dendrogramme (figures 1.5b), il apparaît que les DEGs seuls ne sont pas assez discriminant pour stratifier correctement les patients.

1.4 Conclusion et approche proposée

La recherche d'éléments spécifiques à un phénotype pathologique d'intérêt reste un challenge, particulièrement dans le cas de maladies complexes comme le SLE. En effet, il s'agit d'une

maladie hétérogène et diffuse. Chaque patient peut avoir des symptômes différents et donc des dérégulations différentes. Le fait de réaliser des analyses statistiques comme le calcul de DEGs sur toute une population malade ne permet pas de prendre en compte cette hétérogénéité et les stratifications cliniques habituelles ne semblent pas efficaces.

De plus, l'analyse des données d'expression n'est pas suffisamment discrimante dans ce type de maladie. En effet, comme les évaluations l'ont montré dans la section précédente, les DEGs seuls ne permettent pas d'obtenir une signature de diagnostique qui permette de classer correctement les échantillons. Il est donc nécessaire de déterminer une signature qui puisse tout d'abord prendre en compte l'aspect hétérogène de la pathologie et qui ne se limite pas non plus aux données d'expression.

Afin de répondre aux limites présentées précédemment, l'objectif de la thèse a été de mettre en place une méthode qui permet tout d'abord de prendre en compte l'hétérogénéité de la population. Pour cela, l'hypothèse avancée est de ne plus appliquer d'analyse populationnelle mais de privilégier une analyse à l'échelle du patient. Cela sous-entend que les modèles statistiques classiques ne seraient donc plus applicables dans ce contexte particulier.

La seconde limite énoncée est l'utilisation d'un seul type de données dans la recherche de motif de signature. Dans les pathologies, lorsque l'organisme se dérégule, les cellules vont avoir un comportement dysfonctionnel. Elles peuvent produire des protéines différentes par rapport à celles d'un organisme ayant un comportement normal. Cependant, cette dérégulation pourrait être associée à la cause de la maladie tout comme à ses nombreuses conséquences. Les données d'expression de gènes sont des données de type observationnel. A elles seules, elles ne permettent pas de déterminer si les DEGs sont des causes de la maladie ou bien les conséquences du dérèglement. Il est donc nécessaire d'ajouter des informations supplémentaires à cette première analyse, comme par exemple d'autres données omiques disponibles.

Dans le cas de la thèse, des données de génotypage sont disponibles. Le génotypage est l'analyse permettant d'identifier des mutations chez les individus. Ces mutations peuvent être caractérisées par un seul changement de nucléotides, on parle alors de SNP, ou par l'insertion ou la délétion de plusieurs nucléotides. Ces modifications peuvent entraîner un changement dans l'expression du gène ou dans sa fonction et peuvent être la cause de dérégulations conduisant à une pathologie. C'est le cas des maladies dites monogéniques, comme la mucoviscidose qui est due à des mutations dans le gène *CFTR*, induisant une perte de fonction de la protéine [54]. D'autres pathologies comme le cancer du colon, résultent d'une accumulation de mutations au cours de la progression tumorale [55]. Il est donc courant que la recherche de causalité d'une pathologie commence par l'analyse des mutations chez les patients.

Le cancer est cependant un cas particulier puisque la prolifération anarchique des cellules

tumorales entraîne de l'instabilité génomique. Des dommages de l'ADN font donc aussi partis des conséquences de la maladie, même s'ils ont tendance à toucher les voies de régulation de réparation de l'ADN.

Dans le contexte de la thèse, l'utilisation de ce type de données permettrait d'ajouter un aspect causal dans la recherche de la signature. En combinant les données de génotypage et transcriptomique, il est possible d'obtenir des informations de cause et de conséquences biologiques directes sur les gènes. Pour cela, les technologies du Web Sémantique seront utilisées puisqu'il s'agit d'outils qui permettent de relier des données hétérogènes dans un même modèle.

L'intégration de connaissances pourrait aussi raffiner cette signature en rajoutant cet aspect causal. En effet, les réseaux de régulation sont des connaissances qui permettent de relier les gènes selon les effets des uns par rapport aux autres (activation, inhibition...). Ainsi, les dépendances de régulation entre les gènes seraient une information supplémentaire qui pourrait apparaître dans ces signatures.

L'objectif général de cette thèse est donc de définir une signature causale à l'échelle du patient et en prenant en compte plusieurs types d'informations : plusieurs couches de données omiques ou des connaissances comme des réseaux de régulations. Il sera possible *in fine*, de proposer une nouvelle stratification des patients. En effet, les patients seraient regroupés lorsqu'ils présentent une signature causale similaire.

Chapitre 2

Intégration des données multiomiques à l'échelle du patient par les technologies du Web Sémantique

Les analyses statistiques classiques sur les données omiques permettent d'identifier un signal fort, significatif, qui est relatif à une population homogène d'intérêt. Ce signal va caractériser la population d'intérêt, être sa signature moléculaire.

L'application de ces méthodes dans le chapitre précédent montre qu'il existe des limites sur l'utilisation d'une analyse populationnelle. Tout d'abord, comparer une population par rapport à une autre sous-entend que chaque population est relativement homogène, que les patients sont malades via les mêmes mécanismes moléculaires. Hors, le SLE est décrit comme très hétérogène dans les symptômes, mais aussi dans ses données d'expression.

Ensuite, pour avoir une puissance statistique suffisante, la taille totale des échantillons doit être grande et le nombre d'individus similaire entre les deux populations. Dans le cadre d'une étude clinique comme celle du SLE, le nombre d'individus malades est bien supérieur à celui des individus contrôles.

Enfin, les signatures moléculaires sont généralement constituées d'un seul type de données omiques. Lorsque plusieurs couches omiques sont disponibles, la signature est une union des résultats calculés de manière indépendante.

Il apparaît donc nécessaire de **définir une méthode qui calcule des signatures moléculaires en prenant en compte deux points importants**. Tout d'abord, **utiliser le maximum de données à disposition dans les études multi-omiques**. L'étude clinique SLE contient des données génomiques et transcriptomiques. Ces deux types peuvent être connectés grâce à l'influence que peuvent avoir les SNPs sur la variation d'expression des gènes. Ensuite, chacune de ces données est mesurée pour chaque individu. Il est donc intéressant de **générer des hypothèses moléculaires qui seraient spécifiques à un sous-groupe de patients**

et non pas à toute la population.

Une autre problématique rencontrée est que les données en tant que telles sont d'une part, très hétérogènes dans leur contenu. En effet, les données cliniques comme les données génomiques sont un mélange de données numériques et de chaînes de caractères. D'autre part, elles sont aussi très différentes dans leur format. Les données d'expression sont représentées en matrice, celles de génotypage sont sous format VCF et un tableau regroupe les données cliniques.

Ce chapitre présente la méthode développée au cours de ma thèse et qui répond à ce besoin de prise en compte des caractéristiques de chaque patient mais aussi de connexion entre les couches omiques.

La section 2.1 va introduire l'état de l'art sur ces deux aspects. La première sous-section décrira des méthodes qui favorisent une analyse centrée sur le patient et non pas à l'échelle d'une population d'intérêt. Les principales limites de ces méthodes sont le besoin de connaissances comme des réseaux de régulation pour la personnalisation des données et l'utilisation de données continues uniquement. La deuxième sous-section s'intéresse aux approches existantes qui réalisent des analyses multi-omiques afin d'obtenir des motifs complexes d'éléments biologiques qui sont caractéristiques d'une population. Enfin la troisième sous-section présentera la base de données TCGA comme structure d'intégration de données multi-omiques et à l'échelle du patient.

En réponse aux différentes limites soulevées dans la première section, la première contribution de ce chapitre est la proposition d'un modèle d'intégration transomique en section 2.2. Il s'agit d'une structuration générale des données omiques issues d'une étude clinique répondant au besoin d'analyse des données omiques. Cette structure est présentée sous forme de schéma RDF organisant les relations entre les couches omiques.

Une fois la structure définie, l'alimentation du modèle sera présentée à la section 2.3. Cette alimentation passe par une transformation des données brutes d'une étude clinique afin d'intégrer le besoin de mise à l'échelle du patient. Un schéma récapitulatif de l'intégration des données, à partir des données brutes jusqu'au triplestore final, sera présenté en section 2.4.

La méthode décrite sera appliquée sur une étude clinique proche de celle de SANOFI mais déjà publiée. Les sous-sections contenues dans la section 2.5 présenteront tout d'abord l'étude publiée de Panousis et al., puis appliqueront les étapes d'intégration des données en une base de données locale.

Des étapes de validation sont nécessaires afin de vérifier l'intégration des données multi-omiques dans une approche centrée sur le patient. Dans la section 2.6, trois requêtes simples de validation seront discutées.

Enfin, plusieurs aspects de la méthode seront discutés en section 2.7.

2.1 Etat de l'art sur l'analyse centrée sur le patient et l'intégration multi-omique

2.1.1 Les méthodes centrées sur l'individu, l'échantillon

Les analyses statistiques classiques comme décrites dans le chapitre 1 ne permettent pas une caractérisation de l'hétérogénéité des phénotypes dans une pathologie complexe. Afin de pallier cette limite, les recherches tendent à se focaliser sur des groupes plus petits de patients qui présenteraient des similitudes moléculaires ou biologiques, mais cela se fait au détriment de la puissance statistique.

Ainsi en clinique, la médecine de précision qui tend à personnaliser les traitements, s'est grandement développée ces dernières années. Elle permet de prendre en compte les caractéristiques spécifiques de chaque patient dans sa prise en charge clinique. Certaines méthodes cherchent donc à personnaliser les analyses en les rendant spécifiques à un échantillon, un individu, un patient (Tableau 2.1). Selon le contexte des données biologiques, la question peut être de développer des marqueurs de diagnostic d'une pathologie ou de réponse à un traitement particulier.

Types d'approche	Limites	Outils
Méthodes basées sur les données		
Réseau de corrélation d'expression des gènes	Pas de données discrètes	SSN [56]-[58], ssNPA [59]
Méthodes basées sur les connaissances		
Contextualisation des données sur un réseau	Noeuds du réseau définis a priori Nécessite des interactions formelles	PROFILE [60], CNORode [61]
Personnalisation de l'enrichissement des voies de régulations	Ensembles de gènes définis a priori	ssGSEA [62], Pathifier [63]

TABLE 2.1 – **Tableau récapitulatif d'une liste non exhaustive des méthodes de personnalisation des données et de leur limite**

Un premier ensemble d'outils va utiliser des méthodes statistiques afin de calculer des réseaux globaux à partir d'expression de gènes d'une population de référence. La deuxième étape consiste à comparer ce graphe de référence à des graphes personnalisés pour chaque échantillon d'intérêt. Liu *et al.* ont développé une méthode permettant de construire un réseau spécifique à chaque individu en se basant sur leurs données d'expression [56], appelé SSN. La première étape consiste à calculer un premier réseau de référence en calculant les coefficients de corrélation de Pearson (CCP) entre gènes via leur expression dans un groupe d'individus contrôles. Puis, pour chaque individu du groupe d'intérêt, les valeurs d'expressions de ses gènes sont ajoutés à la matrice de référence et un nouveau réseau de corrélation, un *réseau perturbé* est calculé. Le SSN est défini comme étant le réseau différentiel obtenu entre le réseau de contrôle et le réseau

perturbé. Cette méthodologie a été ensuite réutilisée dans d'autres publications [57], [58]. ssNPA (*single sample Network Perturbation Assessment)* est une méthode récente qui calcule un graphe dirigé de référence en utilisant l'approche de maximisation du Critère d'Information Bayesien (BIC) [59]. Puis de la même manière que pour le SSN, un score de déviation est calculé entre le réseau perturbé d'un échantillon et le réseau de référence.

Les méthodes précédentes calculent des réseaux dont les arcs ne représentent pas des régulations entre les noeuds mais des corrélations entre les expressions. D'autres méthodes de personnalisation se basent sur des connaissances biologiques. Elles nécessitent d'extraire un ensemble de règles régissant les activations et inhibitions entre les éléments biologiques. Il peut s'agir de réseaux de régulation, de signalisation ou d'interaction protéine-protéine selon le type de données biologiques disponibles.

Certains outils sont développés afin de personnaliser des modèles logiques. Par exemple, PROFILE va combiner différentes données omiques avec des réseaux de régulation de gènes [60]. Il utilise des données discrètes telles que l'annotation de mutations somatiques (perte ou gain de fonction), ou de CNV (amplification ou délétion homozygote) et des données continues comme l'expression des gènes, qui va nécessiter une binarisation. Dans un contexte de personnalisation de thérapies pour des patients atteints de cancer, CNORode est une approche permettant d'identifier des biomarqueurs de réponse à des traitements qui seraient spécifiques à des lignées cellulaires de cancer du colon [61]. Un réseau de connaissances est extrait de données de littérature (PKN) sous forme d'équations différentielles (ODE). Il est ensuite personnalisé pour chaque lignée cellulaire en utilisant les données de phosphoprotéomique.

Ces méthodes sont appliquées sur des données de cancer. Cela s'explique par le fait que les bases de données contiennent des réseaux biologiques mieux annotés dans le contexte du cancer (ACSN [64], KEGG, Reactome).

D'autres méthodes vont elles aussi utiliser les informations sur les voies de régulation des bases de données. Ces informations ne sont pas extraites sous forme de réseau mais servent à quantifier l'activation de certaines voies à l'échelle des patients. Pour ssGSEA, l'information est représentée par les ensembles de gènes fonctionnels dans l'analyse d'enrichissement pour chaque échantillon [62]. Quant à Pathifier, il calcule un score pour chaque individu et chaque pathway pour estimer la déviation par rapport à une valeur normale [63].

Les méthodes de personnalisation des données vont soit utiliser uniquement des données continues comme les données d'expression, soit nécessiter des connaissances préalables avec les PKN par exemple, pour mettre en correspondance les données disponibles avec les noeuds des réseaux. Dans des analyses exploratoires, comme c'est le cas dans cette thèse, le fait d'utiliser des PKN pourrait biaiser les résultats. En effet, l'extraction des réseaux de connaissances est une étape difficile et seuls les réseaux les mieux annotés sont les plus informatifs. Ainsi, ces

méthodes ne font que contextualiser les données multi-omiques, elles ne les intègrent pas.

Comme expliqué dans le premier chapitre, la recherche de biomarqueurs se fait généralement à un seul niveau moléculaire. Lorsque des données multi-omiques sont disponibles, leurs analyses se font indépendamment les unes des autres. Cela ne prend pas en compte les relations directes entre les éléments biologiques lorsque les données sont appariées. Afin de répondre à ce besoin d'intégration, plusieurs approches ont été développées pour analyser de manière simultanée les couches omiques. Elles utilisent des aspects mathématiques, de machine learning ou encore des ressources de connaissance comme des réseaux biologiques.

2.1.2 Les approches intégratives sur les données multi-omiques

Plusieurs méthodes statistiques ont été développées depuis quelques années pour intégrer les données multi-omiques afin d'identifier des facteurs biologiques associés à des phénotypes particuliers (prédiction de pathologies, stratification des patients) (Tableau 2.2).

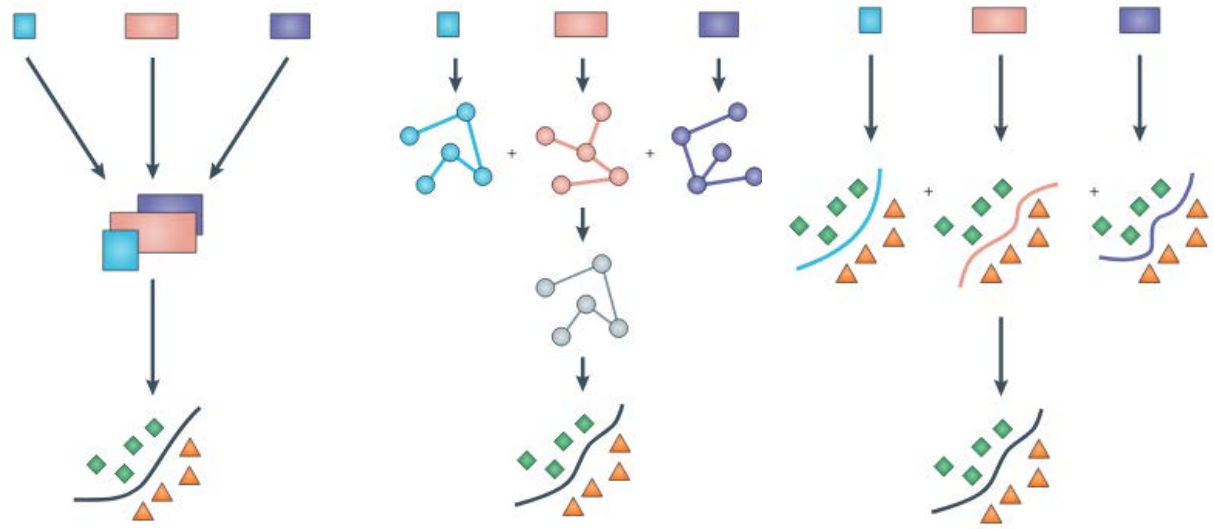
Types d'approche	Limites	Outils
Méthodes mathématiques et de machine learning		
Intégration précoce (Early integration)		
Concaténation des matrices	Etapes de réduction des données Déterminer à l'avance l'intégration Analyse sur population	ATHENA [65] Modèle bayésien [66] Modèle de Lasso [67]
Intégration tardive (Late integration)		
Représentation des couches sous forme de graphes	Perte de l'information d'interaction entre les couches Analyse sur population	Graph-based SSL [68] SVM [69]
Représentation des couches sous forme de modèles statistiques	Analyse sur population	ATHENA [65] MOLI [70]
Approches multivariées (PLS, CCA)	Factorisation des matrices Uniquement des données continues Analyse sur population	DIABLO [71]
Méthodes basées sur les connaissances		
Contextualisation des données sur un pathway	Uniquement des données continues Analyse sur population	*footprint* [72]

TABLE 2.2 – **Tableau récapitulatif d'une liste non exhaustive des méthodes d'intégration de données multi-omics et de leur limite** (inspiré de Ritchie *et al* [73])

Une classification des différentes méthodes d'intégration a été proposée par Ritchie *et al* [73]. Elles sont tout d'abord séparées en deux groupes selon l'ordre d'intégration des données. Une

analyse séquentielle des données est appelée analyse en plusieurs étapes. On y retrouve par exemple les analyses de variation génomique qui se déroule en 3 étapes (méthode en triangle). À partir d'un ensemble de SNPs, la première étape consiste à ne garder que ceux qui sont significativement associés au phénotype d'intérêt. Parmi cette sélection, la deuxième étape va tester les SNPs sur leur association avec une autre couche de données (identification des eQTLs par exemple). Enfin, la dernière étape va calculer la corrélation entre cette dernière couche omique et le phénotype d'intérêt. Cette approche présente une limite lorsque ce que l'on cherche est le résultat d'une combinaison entre différentes couches omiques de manière simultanée.

Les analyses de type méta-dimensionnelle vont permettre d'analyser de manière simultanée les différentes données omiques. Trois catégories ont été définies par Ritchie *et al* [73] et une autre par Tini *et al* [74].



(a) Analyse par concaténation (b) Analyse par transformation (c) Analyse par modèle

FIGURE 2.1 – **Catégorisation des méthodes d'analyse méta-dimensionnelle (Figure issue de [73])**

Les méthodes basées sur la concaténation des données vont permettre de combiner plusieurs matrices de données en une seule ("*early integration*") afin d'y appliquer une seule analyse statistique i.e, Multiple Factor Analysis (MFA), (Figure 2.1a). C'est le cas d'outils qui utilisent le machine learning (réseau de neurones) (ATHENA [65], MOLI [70]), les réseaux bayésiens [66] ou les modèles de LASSO [67]. Comme le souligne Ritchie *et al*, il est nécessaire de bien déterminer à l'avance la façon d'intégrer toutes ces données en une seule matrice. De plus, une étape de réduction de données est nécessaire car la matrice pourrait être trop grande pour être analysée.

D'autres méthodes vont avoir une étape intermédiaire avant de combiner les différentes données ("*late integration*").

Une représentation de chaque donnée omique sous forme de graphe ou de matrice à noyaux est spécifique des méthodes basées sur la transformation (Figure 2.1b) [75]. Dans le cas de graphes par exemple, les noeuds représentent les échantillons ou individus et les arcs sont les possibles relations entre les noeuds et sont caractérisés par chaque donnée omique. Toutes ces représentations seront ensuite fusionnées en un seul objet pour continuer l'analyse. Cela permet de garder les caractéristiques spécifiques apportées par chaque couche omique. En revanche, le risque est la perte de l'information d'interaction entre ces différentes couches omiques.

D'autres approches vont appliquer des approches statistiques sur chaque couche omique comme étape intermédiaire et sont appelées méthodes basées sur les modèles (Figure 2.1c). Ces éléments sont ensuite combinés en un seul modèle final qui sera lui aussi analysé. C'est le cas de MOLI [70] qui génère un profil pour chaque type de données *via* un réseau de neurones profond.

Plus tard, Tini et al. [74] ont défini une nouvelle catégorie d'intégration, les méthodes multivariées, qui utilisent les régressions partielles par les moindres carrés (PLS) et les analyses de corrélation canonique (CCA). C'est le cas de l'outil DIABLO [71] qui pré-analyse chaque donnée omique en la factorisant puis en maximisant la covariance entre les variables latentes (PLS).

D'autres méthodes intégratives ne se basent pas uniquement sur des modèles mathématiques ou de machine learning seuls. Elles utilisent des ressources de connaissances comme des bases de données ou des réseaux de régulation et de signalisation afin d'obtenir des informations supplémentaires sur les interactions biologiques entre les couches. Dugourd *et al.* [72] ont défini le concept de *footprint* ou empreinte, comme étant un ensemble d'éléments liés par des pathways ou des processus d'intérêt. Les connaissances ajoutées aux données proviennent de bases de données et peuvent définir des relations indirectes entre les molécules, comme leur association à des processus biologiques (Gene Ontology). Les molécules peuvent aussi être liées de manière directe, par des relations de régulation comme les enzymes avec leur substrats, les facteurs de transcriptions avec leur gène. Ces relations directes peuvent aussi être extraites à plus large échelle en utilisant les réseaux de régulation, de signalisation ou d'interaction protéine-protéine issus de KEGG, Reactome, Omnipath ou PathwayCommons. Cependant ce type d'analyses semble applicable sur des données multi-omiques continues. En effet, la mesure des empreintes, que cela concerne les facteurs de transcriptions, les enzymes ou les voies de régulation, utilise le changement d'abondance entre les deux éléments (source et cible) pour mesurer l'activité.

En conclusion, selon les types de données omiques, plusieurs méthodes sont disponibles qui utilisent différentes technologies pour identifier des signaux similaires entre les différentes couches omiques. Ces signaux ont pour but d'être réutilisés en tant que critères de classification pour une autre population. Certaines de ces méthodes nécessitent des données continues comme pour les approches basées sur les connaissances, ou ne prennent pas en compte les liens biologiques entre les différentes couches avec les analyses méta-dimensionnelles. Un autre point limitant est qu'elles comparent des populations entre elles et ne passent donc pas à une échelle plus fine,

celle du patient.

2.1.3 Le concept de transomique

Depuis le milieu des années 2010, le terme « transomique » est de plus en plus cité dans les études comme étant un croisement de plusieurs données omiques. À la différence du concept de multi-omiques qui correspond à un regroupement sans liens entre des données omiques de différents niveaux, les analyses transomiques sont des méthodes qui combinent les éléments de plusieurs couches omiques en conservant les relations entre ces couches [76]. Il ne s'agit pas d'associer les signaux les plus forts entre eux mais d'identifier des marqueurs, pour lesquels des éléments d'une première couche ont de l'influence sur une seconde. Plusieurs études appliquent l'approche transomique dans l'identification des mécanismes métaboliques. Des données de métabolomique ont été reliées avec des données d'expression de gènes et de protéines afin d'expliquer la réponse d'*E. Coli* lors de perturbations génétiques ou environnementales [77]. L'abondance de lipides dans le foie a été analysée chez des souris, en combinant des données de protéomiques et de lipidomiques avec de la génomique [78] mais aussi des données de transcriptomique et de métabolomique [79]. Des données de miRNA ont été connectées avec du métabolome pour identifier les mécanismes moléculaires impliqués dans la perte de poids suite à la mise en place d'un bypass gastrique chirurgical [80].

Yugo et al., ont résumé en 2016, plusieurs technologies pour connecter les données multi-omiques en se basant sur les réseaux biologiques [81] :

- régulation métabolique : les données métaboliques sont au centre de l'analyse et leurs mécanismes sont expliqués par d'autres couches ;
- régulation transcriptionnelle : les données des facteurs de transcription (protéome, phosphoprotéome) sont connectées aux données d'expression de leurs gènes cibles ;
- relation kinase-substrat : l'un des cas où une couche est privilégiée, celle du phosphoprotéome qui peut être expliquée par des phénomènes de cascade d'activation d'autres couches omiques ;
- interaction protéine-protéine : comme la régulation précédente, selon les types d'éléments liés, l'interaction peut expliquer d'autres niveaux omiques ;
- régulation allostérique : la connexion entre les métabolites (métabolome) ayant un rôle d'activateurs ou d'inhibiteurs sur des enzymes métaboliques (protéome).

Les auteurs ont défini trois concepts importants dans les analyses transomiques par réseau. Le premier est la connaissance de l'ensemble des interactions possibles entre les éléments biologiques, représentée sous forme de réseau. Le second est le flux de signal statique qui correspond au sens des interactions au sein de réseau (source-cible). Enfin, le troisième concept est le flux de signal dynamique qui se sert de l'aspect quantitatif des données omiques. Par exemple, les études temporelles sont un bon moyen de définir ce type de flux, puisque les valeurs d'abondance des

molécules varient au cours du temps.

Même si l'article présente plusieurs aspects d'une approche utilisant les réseaux, les auteurs expliquent qu'une autre approche est possible. Il s'agit d'analyse transomique basée sur les données qui infère les interactions entre les éléments biologiques par des approches statistiques, de corrélation. Cependant, les auteurs soulignent le fait que cette méthode ne reflète pas directement les interactions biochimiques qui seraient retrouvées grâce aux connaissances.

En se basant sur la définition du concept " transomique ", le modèle proposé en section 2.2 est une approche qui permet de croiser des données multi-omiques. Il peut donc lui aussi être défini comme étant un modèle transomique.

2.1.4 The Cancer Genome Atlas : Un modèle transomique et centré sur le patient ?

La base de données TCGA - The Cancer Genome Atlas (TCGA) est une base de données qui stocke de très nombreuses données multi-omiques (génomique, épigenomique, transcriptomique et protéomique) mesurées sur près de 33 cancers [1]. Le portail permet d'avoir accès aux informations génomiques (mutation, CNV) selon certains critères définis (impact de la mutation, type de gènes, caractéristiques du patient etc...) ainsi qu'aux profils d'expression. Des outils de visualisation ont été créés pour travailler sur les autres couches omiques. Cependant, ils sont généralement limités à l'analyse d'un seul niveau omique. C'est le cas de The Cancer Proteome Atlas Portal (TCPA) [2] pour la protéomique, FASMIC [3] pour l'impact fonctionnel des mutations ou TANRIC [4] pour les ARN longs non-codants. D'autres intègrent plusieurs couches mais la recherche est très limitée. Une liste de gènes doit être préalablement identifiée, un algorithme est déjà en place pour répondre à une problématique précise. SurvNet [5] corrèle par exemple la survie avec un réseau de biomarqueurs omiques. Il n'est pas possible d'avoir une vision d'ensemble des données stockées ni d'appliquer des requêtes exploratoires, spécifiques à un projet.

Les structures du web sémantique appliquées au TCGA - Certaines études ont proposé un schéma d'intégration de ces données en utilisant les technologiques du web sémantique. Des données hétérogènes peuvent donc être intégrées dans un même modèle et des requêtes SPARQL sont réalisées pour en extraire l'information.

Une première méthode publiée en 2010 a assigné les éléments contenus dans TCGA aux entités présentes dans le modèle du Simple Sloppy Semantic Database (S3DB) (Figure 2.2) [82]. Il est nécessaire d'adapter les données selon un schéma d'intégration général, déjà défini et appli-

1. https ://portal.gdc.cancer.gov/
2. http ://www.tcpaportal.org/
3. http ://bioinformatics.mdanderson.org/main/FASMIC
4. http ://bioinformatics.mdanderson.org/main/TANRIC :Overview
5. http ://bioinformatics.mdanderson.org/main/SurvNet

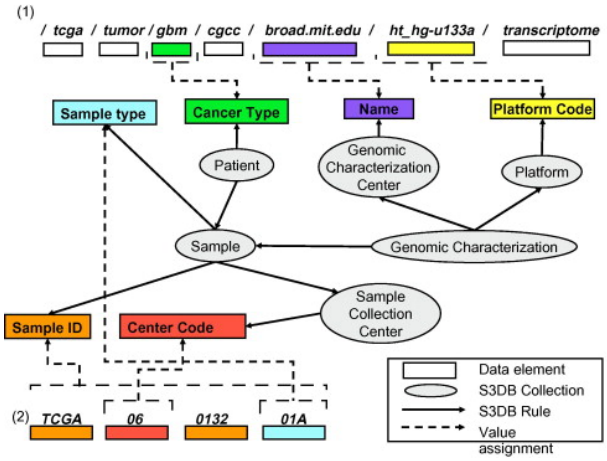


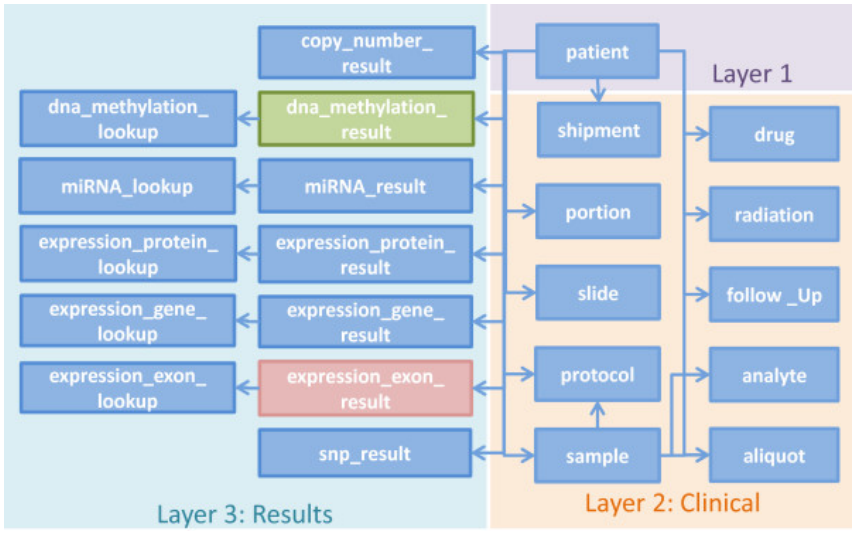
FIGURE 2.2 – Schéma RDF de l'association des données de TCGA avec les entités de S3DB (issu de [82])

cable à n'importe quel domaine. Cette étape d'adaptation nécessite des connaissances préalables en Web Sémantique.

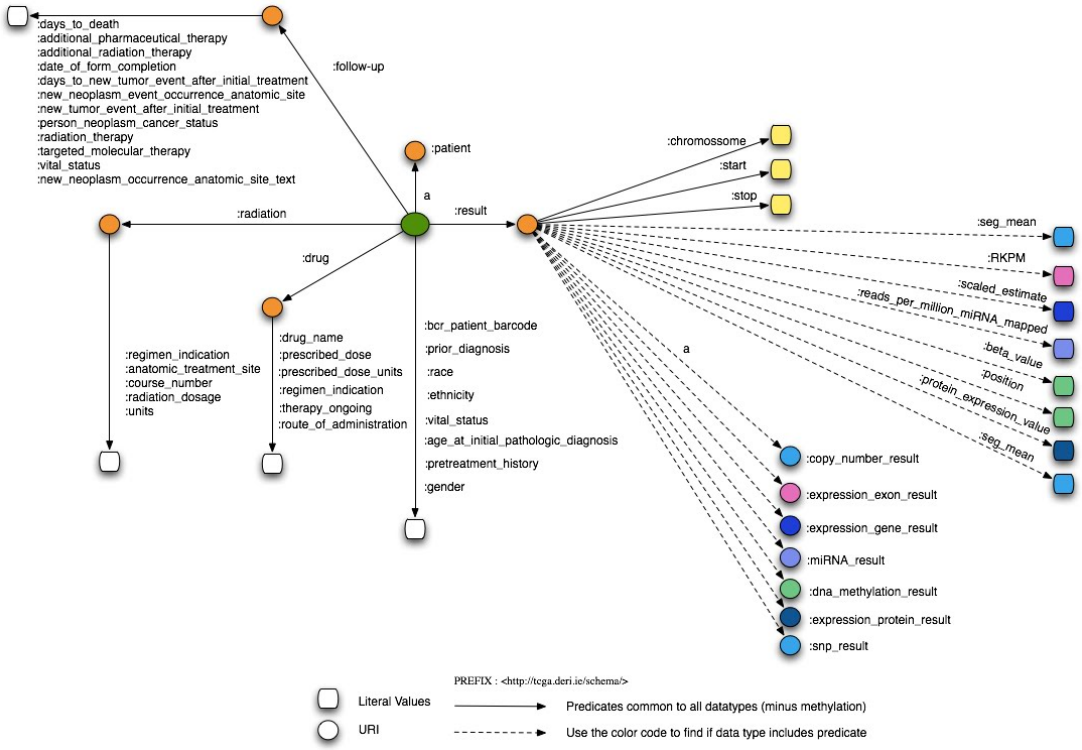
TopFed quant à lui, transforme les données de TCGA en différents endpoints [83]. Pour accéder aux données, des requêtes fédérées, qui sont des requêtes SPARQL permettant d'extraire des informations sur différents endpoints, sont créées. Les données sont structurées en 3 couches : (1) pour les données de patients, (2) pour les informations cliniques, (3) pour les résultats biologiques (Figure 2.3a). Cette dernière couche est celle contenant le plus d'informations. Les auteurs ont attribués 17 endpoints SPARQL à ces données, séparés en 3 groupes : (a) 9 endpoints "green" pour les résultats de méthylation, (b) 6 endpoints "pink" pour les résultats d'expression d'exon et (c) 2 endpoints "blue" pour le reste des données (Figure 2.3b). En se basant sur cette même méthode de requête fédérées, ils ont pu connecter cette base de données RDF à d'autres bases de données comme HGNC [6], OMIM [7] ou Homologene [8].

Les limites de TCGA et de ces modèles - TCGA est une grande base de données stockant des données de différentes couches omiques. Elle permet principalement d'accéder aux différentes archives collectées sur le cancer. Il existe des outils d'analyse de cette base, présents sur le site, qui permettent de comparer des cohortes sélectionnées selon certains critères (type de cancer, genre des individus etc...). D'autres servent à l'exploration des données mais aucun ne permet de faire une sélection fine des éléments biologiques, par l'utilisation de requêtes complexes par exemple. Une autre limite est l'impossibilité de confronter des jeux de données générés localement à l'ensemble des données stockées dans TCGA.

6. http ://hgnc.bio2rdf.org/sparql
7. http ://omim.bio2rdf.org/sparql
8. http ://homologene.bio2rdf.org/sparql



(a) Représentation des couches de données de TopFed (issu de [83])



(b) Schéma RDF de TopFed (issu de [83])

FIGURE 2.3 – Représentations de la base de données TCGA par l'outil TopFed

Les deux outils présentés dans la section précédente, grâce à la technologie du Web Sémantique permettrait de réaliser l'interrogation de cette base de données.

Plusieurs limites apparaissent avec les outils TopFed et S3DB/TCGA. La première, commune aux deux, est l'impossibilité d'ajouter les données d'un autre projet qui ne serait pas référencé dans TCGA. Il pourrait être intéressant de pouvoir enrichir la base de connaissance avec les données d'autres études, même non encore publiées.

Une deuxième limite, déjà évoquée précédemment, est la nécessité de maîtriser la structuration RDF afin de découper les données pour associer les entités de S3DB avec les éléments biologiques des études. Cela peut être un frein à l'utilisation de cette technologie malgré le besoin d'intégrer les données.

TopFed permet de sélectionner une liste de gènes pour un patient selon de valeurs brutes d'expression de gènes ou de protéine (RPKM, estimation réduite etc...). En revanche, il n'est pas possible d'extraire les gènes qui sont sur ou sous exprimés chez un individu malade par rapport à un individu sain par exemple. De plus, cet outil n'est plus maintenu.

Il n'existe pas à notre connaissance, d'autres bases de données structurées grâce aux technologies du Web Sémantique qui permettent de relier des données multi-omiques issues d'un projet personnel, non stockées dans une plus grande base de données.

Afin de pallier ces limites, la prochaine section décrit un schéma d'intégration RDF applicable pour n'importe quelle étude clinique possédant des données cliniques, de génomique et de transcriptomique.

2.2 Un schéma d'intégration de données associé à une étude clinique

Les technologies du Web Sémantique permettent d'obtenir une structuration des données pour leur exploration. Cette organisation favorise l'intégration d'une grande quantité de données (Big data) qui peuvent être hétérogènes. Elle réalise aussi des connexions entre ces différentes informations en construisant un schéma représentatif de cette intégration. Ce dernier peut être réutilisé pour différents jeux de données avec les mêmes structures.

La première étape dans la génération de ce schéma RDF est de faire un premier tri afin d'identifier les éléments qui sont intéressants à extraire lors de requêtes. Ainsi le choix s'est porté sur les individus et leur caractéristiques (âge, ethnie, symptômes...), les gènes et leurs attributs (chromosome, position, fonction...), les génotypages et les valeurs d'expression de chaque gène pour chaque individu.

Par la suite, les connexions entre les données sont définies. De plus, il s'agit de données appariées, ce qui signifie que pour chaque individu, les informations d'expression de gène et de génotypage sont disponibles. La méthode va donc lier le maximum d'information, d'entités de la manière suivante :

- un individu peut être caractérisé par des informations personnelles (âge, ethnie, sexe) mais aussi cliniques (maladie, symptômes, traitement) :
  Individual *attribut* valeur
- l'expression d'un gène est mesurée pour chaque individu :
  Expression *measuredIn* Individual
- l'expression d'un gène est quantifiée par une valeur d'expression :
  Expression *value* valeur
- l'expression est relative à un gène :
  Expression *concerns* gène
- la valeur de chaque SNP est mesurée par individu :
  SNPValue *measuredIn* Individual
- la valeur de chaque SNP correspond à la présence/absence du snp :
  SNPValue *has/hasNot* SNP
- la valeur de chaque SNP est dépendante du génotype :
  SNPValue *AlleleA/AlleleB* valeur
- un snp est une mutation présente sur un gène :
  SNP *localisedIn* Gene
- un gène peut avoir plusieurs caractéristiques (chromosome, position, brin, début, fin, nom) :
  Gene *has* attribut

Ainsi pour chaque entité (Fig.2.4), il existe un fichier .tsv.

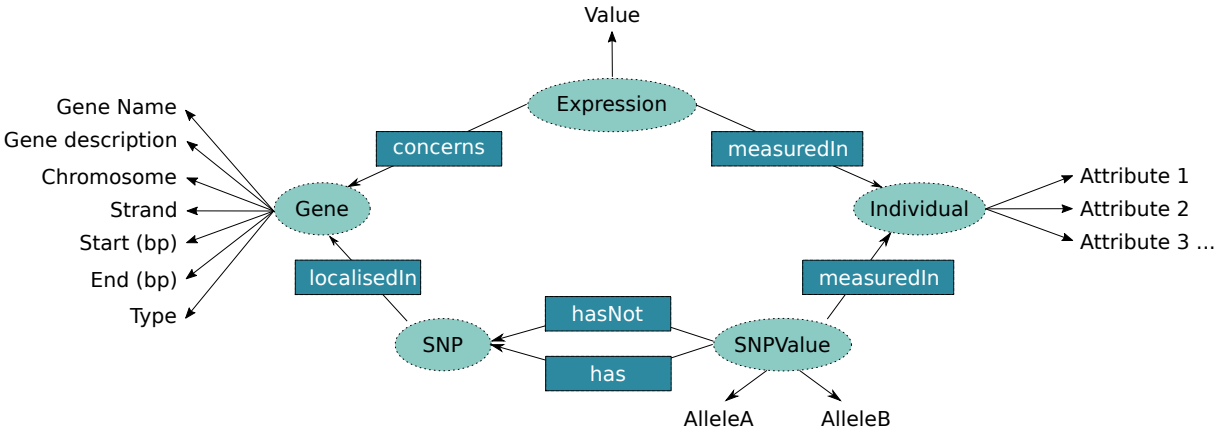


FIGURE 2.4 – **Schema RDF représentant les données** - Les ellipses représentent les classes, les rectangles sont les relations entre entités.

Il est donc possible de relier toutes les informations entre elles et d'avoir accès à ces informations dans un seul et même modèle.

La suite de ce chapitre va présenter l'approche utilisée pour alimenter le modèle d'intégration à l'échelle du patient.

2.3 Alimentation de la structure avec des données multi-omiques centrées sur le patient

L'un des intérêts présentés par ce modèle RDF est de pouvoir réaliser des analyses sur plusieurs couches omiques de manière simultanée. Elle définit un schéma qui va lier l'ensemble des données pour intégrer un jeu de données cliniques, même si celui-ci est hétérogène. L'alimentation de ce modèle nécessite de prendre en compte des données discrètes comme le génotypage des patients ainsi que des données continues telles que les données d'expression.

Les fichiers bruts issus des études biologiques sont aussi hétérogènes dans leur format : des matrices avec des valeurs continues pour les données d'expressions et des fichiers VCF pour les données de génotypage. Il est donc nécessaire d'ajouter une étape de transformation afin d'avoir une valeur par entité et par individu. Cette étape va prendre en compte l'hétérogénéité de la population et conduire à une approche centrée sur le patient.

Dans un premier temps, même si les données de génotypage sont déjà à l'échelle du patient, il est nécessaire de les transformer pour permettre leur intégration dans le modèle transomique. Dans un deuxième temps, les données brutes, lorsqu'elles sont continues, ne permettent pas de réaliser des analyses centrées sur le patient. La variabilité inter-individus doit être prise en

compte dans les calculs. Ces étapes sont donc généralisables aux études présentant les mêmes types de données.

Génomique

Les données disponibles sont des données de génotypage au format VCF. Ce sont des données discrètes qui permettent déjà de comparer les individus ou les gènes entre eux. Elles ne nécessitent pas de grande transformation mais simplement un reformatage (Figure 2.5).

Elles permettent d'identifier la présence ou l'absence d'une liste de SNP définis sur une puce à ADN chez chaque individu. Un SNP peut aussi être caractérisé par la présence en un ou deux exemplaires de la mutation. Les individus peuvent être hétérozygotes ou homozygotes pour chaque SNP. Cette information a été intégrée dans le modèle transomique.

Dans un souci de simplification, il est possible dans le modèle transomique de n'avoir accès qu'à la présence (*has SNP*) ou l'absence (*hasNot SNP*) d'un SNP pour chaque individu.

On obtient donc à partir d'un fichier VCF, un tableur avec l'information de génotypage pour chaque couple individu-SNP avec en ligne les identifiants des individus et en colonne les valeurs du génotypage issu du fichier VCF.

```
##INFO=<ID=PR,Number=0,Type=Flag,Description="Provisional reference allele, may not be based on real reference genome">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
#CHROM  POS ID  REF ALT QUAL    FILTER  INFO    FORMAT  Indiv1  Indiv2  Indiv3  Indiv4  Indiv5  ...
chr 008 rs1:008:C:G C   G   .   .   PR  GT  0/0 0/0 0/0 0/0 0/0 ...
chr 012 rs2:012:C:G C   G   .   .   PR  GT  0/0 0/0 0/0 0/0 0/0 ...
chr 211 rs3:211:T:G G   T   .   .   PR  GT  0/1 0/0 0/1 0/0 0/0 ...
chr 274 rs4:274:A:T T   A   .   .   PR  GT  0/1 0/1 0/0 0/0 0/0 ...
chr 712 rs5:712:T:TTTTC TTTTC   T   .   .   PR  GT  0/1 0/0 0/1 0/1 0/1 |...
```

(a) Fichier VCF de génotypage

SNPID	rs1 :008 :C :G	rs2 :012 :C :G	rs3 :211 :T :G	rs4 :274 :A :T	rs5 :712 :T :TTTTC
Indiv1	0/0	0/0	0/1	0/1	0/1
Indiv2	0/0	0/0	0/0	0/1	0/0
Indiv3	0/0	0/0	0/1	0/0	0/1
Indiv4	0/0	0/0	0/0	0/0	0/1
Indiv5	0/0	0/0	0/0	0/0	0/1

(b) Matrice de résultats discrétisés

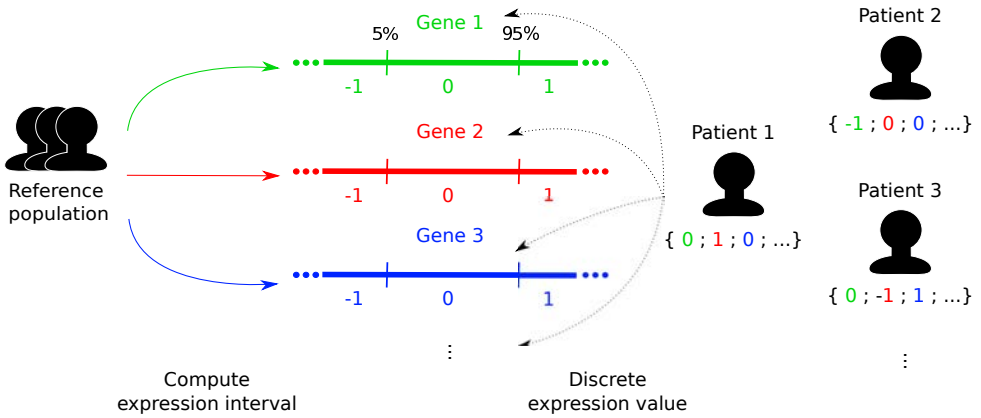
FIGURE 2.5 – **Pré-traitement des données initiales de génomique (fichier VCF) en matrice de génotypage.** (a) Exemple d'un fichier VCF contenant des analyses de génotypage. (b) Exemple de représentation d'un fichier tabulé permettant l'alimentation du modèle transomique.

Transcriptomique :

Les données de transcriptomique, à l'inverse des données de génotypage, sont des données continues. Il s'agit de comptage (count) issu d'une analyse de RNAseq.

Chaque patient a une expression particulière pour chacun de ces gènes. En revanche, le fait de savoir si un individu surexprime ou non un gène nécessiterait de connaître le seuil pour lequel un gène est sur- ou sous-exprimé. Dans les analyses populationnelles et statistiques, on utilise la population de référence ou contrôle pour évaluer si les gènes semblent être de manière significative sur/sous-exprimés dans la population d'intérêt.

Afin de discrétiser les données pour chaque patient, il faut aussi prendre en compte la variabilité d'expression des gènes entre individus. Ainsi la méthode développée utilise les valeurs d'expression de chaque gène de la population contrôle afin d'obtenir un intervalle d'expression. Pour chaque gène de chaque patient, sa valeur est discrétisée de la manière suivante : si la valeur d'expression de ce gène est contenue dans l'intervalle d'expression de la population de référence, alors sa valeur discrétisée est 0. En revanche, si la valeur d'expression est supérieure (ou inférieure) à l'intervalle d'expression, alors la valeur discrétisée est 1 (ou -1) (Figure 2.6).



(a) Méthode de discrétisation par intervalle d'expression

SNPID	Gene1	Gene2	Gene3	Gene4	Gene5
Indiv1	0	0	1	1	1
Indiv2	0	0	0	1	0
Indiv3	0	0	1	0	1
Indiv4	0	0	0	0	1
Indiv5	0	0	0	0	1

(b) Matrice de résultats discrétisés

FIGURE 2.6 – **Traitement des données brutes de transcriptomique (matrice d'expression) en matrice discrétisée. (a) Schéma de la méthode de discrétisation des données d'expression par l'utilisation d'un intervalle d'expression pour chaque gène.**

Pour chaque fichier d'origine, la méthode génère un fichier intermédiaire qui servira d'entrée pour l'étape d'intégration. Le reste des étapes d'intégration suit une même succession, quelque soit l'étude biologique.

2.4 Intégration des données processées dans un endpoint

La figure 2.7 représente les différentes étapes qui sont nécessaires pour passer de données brutes issues d'une étude clinique comme celle de SANOFI à une base de données locale, un modèle d'intégration transomique.

A partir de ces données brutes, on applique l'approche centrée sur le patient pour les données transcriptomiques et on extrait les informations utiles des données génomiques. On obtient donc des matrices avec les éléments biologiques en lignes et les individus en colonnes.

Comme présenté dans l'introduction, les technologies du Web Sémantique répondent à un besoin d'intégration de données complexes et hétérogènes. La représentation des données grâce à ces technologies est souvent appliquée dans les contextes biomédicaux [84], [85]. Cependant, cela nécessite de formater les données dans un cadre particulier. C'est le cas du schéma présenté dans la section 2.2. En effet, cette représentation repose sur la structuration de chaque donnée sous forme de triplets RDF. Un triplet est un ensemble de trois entités de la forme *sujet-prédicat-objet*. Le prédicat va représenter la relation entre le sujet et l'objet. Le sujet et l'objet seront représentés sous forme d'entités. Par exemple, pour les données d'expression, *?Expression* correspond à la variable représentant le sujet, *?Indiv* correspond à la variable représentant l'objet et *ask :measuredIn* est le prédicat.

Ainsi, en reprenant le schéma d'intégration de la thèse, les prédicats sont les arcs entre les noeuds. Les schémas RDF étant des graphes orientés, le noeud source est le sujet et le noeud cible est l'objet.

Pour chaque ligne des fichiers, une relation sera donc créée entre une valeur de la première colonne et une valeur d'une autre colonne sous forme de triplet. La génération d'un triplet issu du fichier de génomique prend la forme suivante : ask :SNPValue ask :hasNot ask :SNP.

Dans le cadre de la thèse, l'ensemble des triplets est généré et stocké dans une base de données locale appelée endpoint par l'outil AskOmics [9]. Il prend en entrée les fichiers tabulés issus de l'étape de transformation des données brutes.

La méthode présentée est une succession d'étapes permettant de passer d'un ensemble de fichiers à une base de données locale. L'intérêt de cette méthode est de pouvoir combiner ces fichiers qui sont hétérogènes et complexes, dans leur contenu et leur forme. Le modèle d'intégration est un modèle transomique, commun pour plusieurs données omiques. De plus, de par la transformation des données d'expression, cette base de données est à l'échelle du patient. Enfin, grâce au schéma général, il peut être appliqué à différentes études, à partir du moment où le type de données disponible est le même que ceux de l'étude clinique de SANOFI.

9. https ://askomics.org/

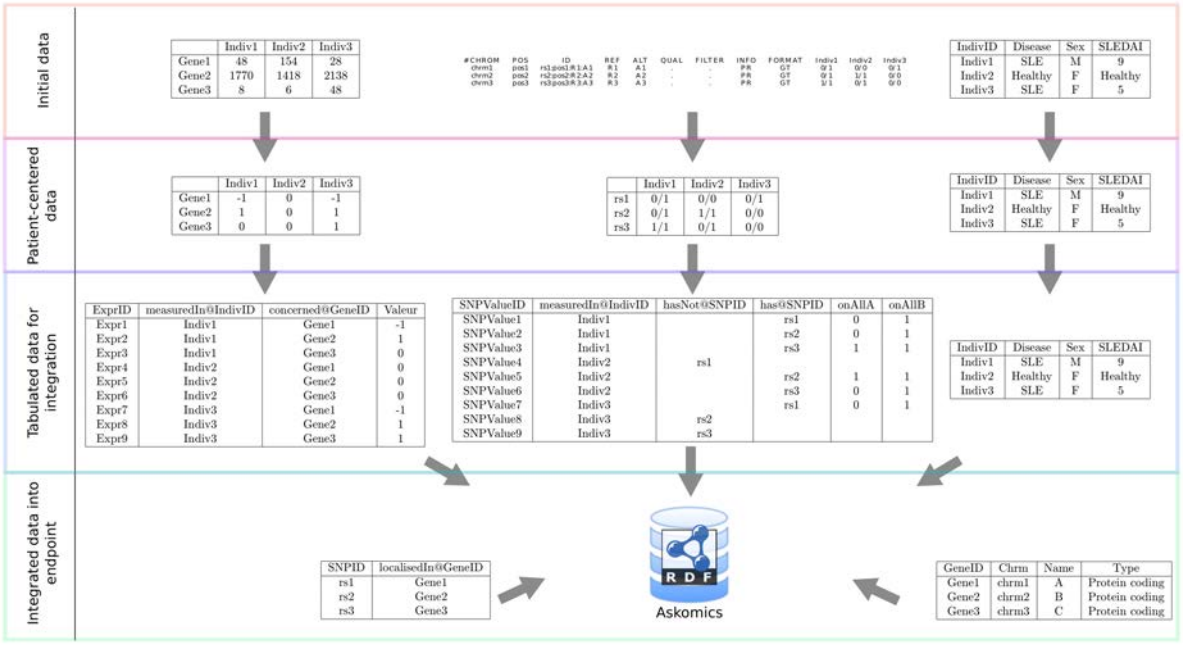


FIGURE 2.7 – **Etapes d'intégration à partir des fichiers initiaux jusqu'au triplestore**

2.5 Modèle final : application aux données de SLE

2.5.1 Présentation de l'étude clinique de Panousis *et al.*

Afin de valider la méthode, il fallait trouver des données publiées, disponibles, pairées, multi-omiques et proches de celle de l'étude du SLE par SANOFI. Il existe de nombreuses études multi-omiques mais le concept de multi-omique englobe de nombreux types (transcriptomique, protéomique etc...). Il a donc été difficile de trouver un jeu de données contenant des données d'expression et des données de génotypage. En effet, dans la majorité des études, les données de génomique concernent des analyses de méthylation ou de CNVs. Les données contenant les SNPs sont sensibles car elles contiennent des informations génétiques sur les patients. En plus de la complexité à trouver des données multi-omiques de type génotype-expression, obtenir des données pairées est tout aussi compliqué. Beaucoup de publications analysent des données omiques issues d'études différentes.

Une étude récente a néanmoins réussi à passer tous ces critères. Il s'agit d'une étude de 2019, portant sur l'identification de signature du SLE en combinant des analyses génétiques et du transcriptome [86]. Le jeu de données concernent 200 individus (142 patients SLE et 58 individus sains) et contient des données de génotypage et de RNA-seq.

Les données multi-omiques ont d'abord été analysées indépendemment les unes des autres. Une liste de DEGs a été identifiée à partir des données d'expression puis enrichie fonction-

nellement par GSEA. En comparant plusieurs types de population, les auteurs ont identifié une signature de susceptibilité, une signature d'activité et une signature de sévérité. Ces trois signatures étant constituées de DEGs.

Les données de génotypage ont été analysées afin d'identifier une causalité génétique. Pour cela, Panousis et al., ont dans un premier temps, filtré les eQTL de GTex avec des données de GWAS du SLE, puis réalisé le même type de filtrage mais avec les DEGs mesurés dans la comparaison de la population SLE par rapport à la population contrôle.

Ce jeu de données correspond au type d'étude pour laquelle la méthode a été développée. De plus, comme discuté dans les sections précédentes, les auteurs utilisent les données multi-omiques de façon indépendante. La prochaine section est donc une application de la méthode sur ce jeu de données.

2.5.2 Intégration des données multi-omiques

Données cliniques - Les données cliniques étant déjà sous forme de fichier tabulé avec les informations d'un individu par ligne, il n'y a pas eu besoin de générer de fichier intermédiaire. Chaque individu est caractérisé par un ensemble de 70 attributs (âge, ethnie, symptômes, concentration cellulaire).

Données d'expression - L'étape de discrétisation a été réalisée pour les données d'expression. Le fichier initial est une matrice d'expression de 41 020 gènes par 200 individus. La discrétisation va générer un fichier tabulé qui contient 8 162 183 relations individu-gène.

Données de génotypage - Les données de génotypage de l'étude correspondaient à des données imputées. Lors du génotypage d'individus à partir de puce à ADN, les SNPs utilisés dans la puce sont appelés Tag SNPs. Il s'agit de SNPs qui sont utilisés comme représentants d'une séquence génomique. L'ensemble des SNPs entre ces Tag SNPs n'est donc pas identifiée. L'imputation est une étape qui va permettre de remplacer les valeurs manquantes de SNPs entre les Tag SNPs à partir d'un panel de génome de référence (ex : les 1000 génomes). Les SNPs imputés seront donc les SNPs qui apparaissent le plus fréquemment ensemble dans la population sachant la valeur des Tag SNPs.

À partir de l'ensemble de ces données, seuls les Tag SNPs qui sont contenus chez au moins un individu ont été gardés. Deux fichiers tabulés ont été créés avec le lien SNP-Gène d'un côté et les valeurs de SNP pour chaque individu à part. Cela permet de sélectionner soit uniquement les individus qui n'ont pas le SNP d'intérêt (*hasNot SNP*) ou ceux qui l'ont (*has SNP)* sans avoir besoin des informations d'hétérozygotie ou d'homozygotie dans un premier temps. Cela représente un total de 6,106 SNPs.

Classe (nombre d'entités)	Attributs
Gene (40,122)	Ensembl ID GeneName GeneDescription Chromosome Start End Type Brin
SNP (9,098)	@Gene
Expression (8,162,781)	Value
SNPValue (1,202,882)	@Individu @SNP AlleleA AlleleB
Individu (200)	Individu ID Maladie Centre Genre Age Ethnie --- Malarrash Photosensitivity Mucosal ulcers Arthritis Serositis Renal disease CNS disease Hematological Immunological ANA No ACR criteria anti.DNA antibodies aPL antibodies APS NPSLE Coexisting autoimmune disease HCQ DMARDs MMF Biological SLE status Physician General constitutional Mucocutaneous Neurological Musculoskeletal Cardiorespiratory Vasculitis skin.GI Renal Haematology Serological activity Clinical SLEDAI Nephritis history Nephritis active --- B.cells.naive B.cells.memory Plasma.cells T.cells.CD8 T.cells.CD4.naive T.cells.CD4.memory.resting T.cells.CD4.memory.activated T.cells.follicular.helper T.cells.regulatory..Tregs. T.cells.gamma.delta NK.cells.resting NK.cells.activated Monocytes Macrophages.M0 Macrophages.M1 Macrophages.M2 Dendritic.cells.resting Dendritic.cells.activated Mast.cells.resting Mast.cells.activated Eosinophils Neutrophils

TABLE 2.3 – **Tableau récapitulatif des classes et attributs contenus dans le modèle commun** Le modèle commun est composé de 5 classes : Gene, SNP, Expression, SNPValue et Individu. Chaque classe contient un nombre d'entités et est défini par un certain nombre d'attributs. Ces attributs peuvent être spécifiques à la classe ou être reliés à d'autres classes (@)

Population des données dans le modèle - Lorsque toutes les données ont été transformées et intégrées dans le modèle transomique, cela représente plus de 9 millions de lignes ou entités (Tableau 2.4). Chaque élément étant représenté sous forme de triplets, il y a un peu plus de 48 millions de triplets (Tableau 2.4). Les informations de chaque classe avec ses différents attributs sont détaillées dans le tableau 2.3.

	Nombre d'éléments
Triplets	48,350,565
Entités	9,415,083
Classes	5

TABLE 2.4 – **Population des données dans le modèle intégré transomique**

La génération des triplets et leur intégration sont réalisées par l'outil Askomics développé au sein du laboratoire. Des étapes de validation sont néanmoins nécessaires afin de s'assurer que les données sont correctement reliées entre elles, comme défini dans le schéma RDF. Pour cela, 3 requêtes SPARQL simples sont testées dans la section suivante, chacune ayant pour objectif de valider un aspect de l'intégration.

2.6 Validation des étapes d'intégration via des requêtes

2.6.1 Intégration des données appariées

Le modèle transomique permet d'avoir accès aux données appariées. Afin de valider cet aspect, une requête a été créée pour lister tous les individus malades qui possèdent au moins une expression de sous-régulation (valeur discrète à -1) et qui possèdent au moins un SNP.

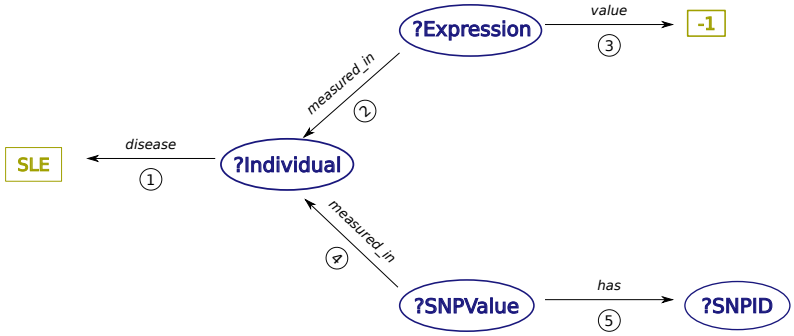
La requête est donc définie par la figure 2.8.

En résultat, on obtient 140 patients SLE sur 142 en 1h09m46s. Cela signifie qu'il y a 2 patients SLE qui soit ne possèdent pas de SNPs, soit ne possèdent pas de gène dont l'expression est sous-régulée, soit un mélange des deux. Ces résultats confirment que le modèle RDF transomique intègre l'aspect appariés des données omiques.

2.6.2 Intégration de données multi-omiques

Le modèle transomique permet d'avoir un modèle commun regroupant toutes les données multi-omiques. Afin de valider cet aspect, une requête a été créée pour lister tous les gènes dont l'expression est sous-régulée (valeur discrète à -1) et qui possèdent un SNP.

La requête est donc définie par la figure 2.9.

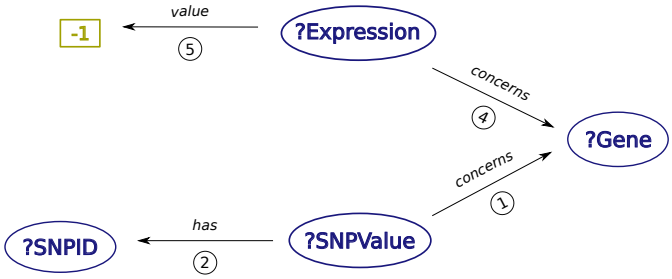


(a) Représentation graphique

```
PREFIX ask: <http://askomics.org/data/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>


SELECT DISTINCT ?INDIV1_Label
WHERE {
    VALUES ?INDIV1_DiseaseCategory { ask:SLE }
    VALUES ?Expression1_Disc { <http://askomics.org/data/-1> }

    ?Expression1 ask:Exp_concern ?Gene1 .
    ?Expression1 rdf:type ask:Expression .
    ?Expression1 ask:Value ?Expression1_Disc .
    ?Expression1 ask:Exp_measurIn ?INDIV1 .

    ?INDIV1 rdf:type ask:INDIV .
    ?INDIV1 ask:Disease ?INDIV1_DiseaseCategory .

    ?SNPValue1 ask:measuredIn ?INDIV1 .
    ?SNPValue1 rdf:type ask:SNPValue .
    ?SNPValue1 ask:has ?SNP1 .
}
```

(b) Requête en langage SPARQL

FIGURE 2.8 – Représentation graphique et corps de la requête SPARQL permettant d'identifier tous les individus SLE dont au moins une expression est à -1 et qui contient au moins une fois un SNP



(a) Représentation graphique

```sparql
PREFIX ask: <http://askomics.org/data/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>


SELECT DISTINCT ?Gene1_Label
WHERE {
    VALUES ?Expression1_Disc { <http://askomics.org/data/-1> }

    ?Expression1 ask:Exp_concern ?Gene1 .
    ?Expression1 rdf:type ask:Expression .
    ?Expression1 ask:Value ?Expression1_Disc .
    ?Expression1 ask:Exp_measurIn ?INDIV1 .

    ?SNPValue1 ask:measuredIn ?INDIV1 .
    ?SNPValue1 rdf:type ask:SNPValue .
    ?SNPValue1 ask:has ?SNP1 .

    ?SNP1 ask:localisedIn ?Gene1 .
    ?SNP1 rdf:type ask:SNP .
}
```

(b) Requête en langage SPARQL

FIGURE 2.9 – Représentation graphique et corps de la requête SPARQL permettant d'identifier tous les gènes dont au moins 1 expression est à -1 et qui contient au moins une fois un SNP

En résultat, on obtient 4 343 gènes sur 41 020 en 3m20s. Cela signifie que seulement 4 343 gènes sont associés à la présence d'un SNP sachant que leur expression est sous-régulée. Les deux couches de données omiques sont donc correctement intégrées dans le modèle.

2.6.3 Intégration de l'approche centrée sur le patient

Une dernière requête combinant les deux précédentes permet de vérifier que l'on peut accéder à des résultats transomiques qui sont centrés à l'échelle du patient.

La requête définie ci-après va lister tous les couples individu-gène où l'individu est diagnostiqué SLE et possède un gène qui est sous-régulé et qui contient un SNP (Figure 2.10).

En résultat, on obtient 19 764 couples patient SLE-gène en 1m6s. L'obtention de ces couples valide l'intégration de données omiques appariées, hétérogènes et centrées sur le patient dans le modèle transomique.

2.7 Discussion

2.7.1 Schéma d'intégration

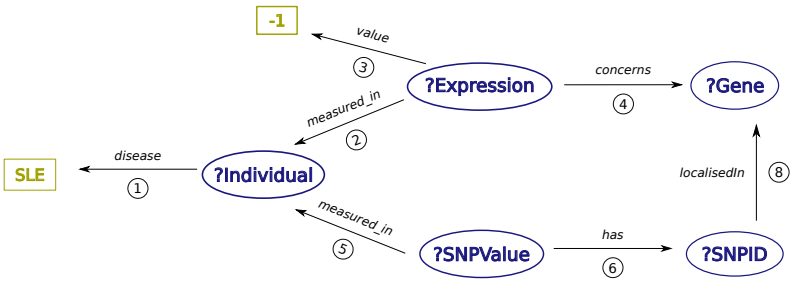
Présence SNP vs Absence/Présence SNP - Un premier tableur a été réalisé où uniquement l'information de présence d'un SNP chez un individu (de manière homozygote ou hétérozygote) était stockée. Cependant, dans la suite de la méthode, cela pose un problème. En effet, en ne gardant que l'information de présence d'un SNP, il est impossible lors de l'élaboration de la requête, d'avoir accès à l'information d'absence d'un SNP chez un individu. Il est apparu nécessaire de modifier ce tableur afin d'avoir accès à l'information de présence mais aussi d'absence de SNP pour chaque individu.

Grâce à cette information d'absence, la requête peut devenir une comparaison entre individus qui possèdent ou non un SNP et donc obtenir l'impact de ce SNP sur une caractéristique d'intérêt.

2.7.2 Sélection des données de génotypage

Etapes de filtrage des SNPs - Plusieurs essais d'intégration ont été réalisés pour les données de génotypage.

Les données imputées représentent environ 3,5 millions de couples SNPs-gène. Lorsqu'on modifie le fichier d'origine en fichier de tabulation pour l'intégration, on obtient un total de 715 872 000 lignes (triplets individu-gene-SNPs). Cependant lors de l'élaboration de la requête qui sera décrite au chapitre 3, le fait d'intégrer toutes les données va entraîner de la redondance dans les résultats. En effet, les SNPs avec un déséquilibre de liaison (LD) fort, apparaîtront ensemble dans les résultats de causalité.



(a) Représentation graphique

```
PREFIX : <http://askomics.org/data/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT DISTINCT ?INDIV1_Label ?Gene1_Label
WHERE {
    VALUES ?INDIV1_DiseaseCategory { ask:SLE }
    VALUES ?Expression1_Disc { <http://askomics.org/data/-1> }

    ?Expression1 ask:Exp_concern ?Gene1 .
    ?Expression1 rdf:type ask:Expression .
    ?Expression1 ask:Value ?Expression1_Disc .
    ?Expression1 ask:Exp_measurIn ?INDIV1 .

    ?INDIV1 rdf:type ask:INDIV .
    ?INDIV1 ask:Disease ?INDIV1_DiseaseCategory .

    ?SNPValue1 ask:measuredIn ?INDIV1 .
    ?SNPValue1 rdf:type ask:SNPValue .
    ?SNPValue1 ask:has ?SNP1 .
    ?SNP1 ask:localisedIn ?Gene1 .
    ?SNP1 rdf:type ask:SNP .
}
```

(b) Requête en langage SPARQL

FIGURE 2.10 – Représentation graphique et corps de la requête SPARQL permettant d'identifier toutes les relations individu-gène dont l'expression du gène est à -1 et qui contient un SNP

Le choix a donc été fait de ne prendre en compte que les SNPs représentatifs des régions, soit les Tags SNPs. Pour cela, on a filtré les SNPs qui étaient référencés dans la documentation du fabricant de la puces à ADN. Ceci a permis de réduire le nombre de SNPs à 261 084 SNPs, pour un total de 28 112 492 lignes. Le temps d'intégration de ce fichier était assez important. Comme la requête du chapitre 3 ne recherche que les SNPs dont la présence chez un individu a un effet, la liste des Tag SNPs a donc été réduite. Pour cela, tous les SNPs qui étaient absents chez tous les individus ont été retirés. La quantité finale de SNPs intégrés est de 6 106.

L'ensemble de ces étapes de filtration des SNPs a permis de ne garder que les SNPs ayant un potentiel informatif et sans avoir de redondance.

2.7.3 Discrétisation des données d'expression

Comparaison à RefBool - L'outil RefBool a été développé pour permettre la discrétisation et donc l'utilisation de données transcriptomiques dans un réseau Booléen [87]. Ce type de réseau est caractérisé par des noeuds qui, dans un contexte biologique, correspondent à des gènes ou des protéines et les arcs signés et dirigés sont les régulations entre ces éléments. Des valeurs discrètes (0 et 1) sont attribuées à chaque noeud afin de représenter l'expression (activation = valeur 1) d'un gène (protéine) par exemple.

Pour cela, la méthode se base sur la distribution empirique de données d'expression de chaque gène issues d'une population de référence. À partir de cette distribution, des seuils sont calculés pour définir les états de chaque gène : actif ou inactif de manière significative et un troisième état associé à une valeur non significative.

En revanche, les auteurs ont montré dans leur étude que la taille de la référence est un facteur important dans leur analyse. En effet, ils ont pu montrer qu'un minimum de 550 échantillons est requis dans la population de référence afin d'approximer la réalité biologique. Il s'agit donc d'une limite importante de l'utilisation de cette méthode avec les jeux de données cliniques. Il est assez rare que le nombre de recrutements dans les études cliniques atteigne un total de 500 individus. Ainsi, obtenir un minimum de 500 échantillons contrôle pour la distribution de référence n'est pas envisageable dans le contexte de la thèse.

Modification des valeurs des bornes - Pour la validation de notre requête, les bornes qui permettent de discrétiser les valeurs ont été fixées à 5 et 95%. Cela représente un faible nombre de valeurs discrétisées. Il est possible de modifier ces bornes et de réaliser un intervalle d'expression plus petit pour chaque gène. Cela pourrait avoir comme conséquence d'augmenter le nombre de gènes dont la valeur serait fixée à 1 ou -1.

Obtenir des valeurs multi-valuées - Notre approche discrétise les expressions entre 3 valeurs possibles : -1, 0 ,1.

Il serait intéressant de réaliser plusieurs seuils dans la discrétisation des expressions et donc d'obtenir un panel plus large de valeurs. Cela permettrait d'identifier les valeurs qui serait plus extrêmes dans les expressions et les discrétiser à +2 dans le cas d'une surexpression ou -2 pour les sous-expressions.

2.8 Conclusion : Génération d'un modèle transomique d'intégration de données cliniques multi-omiques à l'échelle du patient

La problématique était de proposer une méthode prenant en compte deux aspects qui sont des limites dans les identifications classiques de signature : l'utilisation de données multi-omiques et l'analyse à l'échelle du patient.

Dans ce chapitre, un modèle d'intégration a été mis en place afin de pallier ces limites. Il est basé sur des données typiquement mesurées dans beaucoup d'études cliniques : la génomique et la transcriptomique. La méthode intègre des données de couches omiques différentes sous forme d'une base de données RDF accessible via un SPARQL endpoint. Toutes les informations de l'étude clinique dont sont issues les données sont intégrées qu'elles soient continues ou discrètes. De plus, l'apport d'appariements des données est conservé. Enfin, l'alimentation du endpoint est basée sur une approche centrée sur le patient, ainsi les valeurs d'expression et de génotypage sont spécifiques pour chaque patient.

Enfin, cette méthode a été appliquée à un jeu de données publique, proche de celui de l'étude de SANOFI [86]. Le modèle de données transomique centré-patient contient un total de 48 350 565 triplets.

Une fois ce modèle transomique réalisé, il est possible d'avoir accès à l'ensemble des données via l'utilisation de requêtes SPARQL. Ces requêtes vont permettre d'obtenir des informations concernant les gènes et/ou les individus connaissant la valeur d'expression d'un gène ou la présence d'un SNP. Un exemple de requête complexe sera donc expliqué dans le prochain chapitre.

Chapitre 3

Définition, calcul et évaluation des eICTLs candidats

La recherche de marqueurs biologiques spécifiques d'un phénotype passe dans beaucoup d'études par des analyses statistiques classiques, comme évoqué au chapitre 1. En effet, ce type d'approches permet d'identifier un signal fort dans une population homogène.

Dans ce même chapitre, nous avons montré que dans le contexte du SLE, cela induit deux limites majeures. Tout d'abord, l'hétérogénéité de la population ne permet pas de retrouver la stratification classique des échantillons par le calcul des gènes différentiellement exprimés. Ensuite, l'utilisation des données de transcriptomique seules ne permet pas de séparer les marqueurs de cause et de conséquence de la maladie.

Afin de palier à ces limites, une structuration des données a été proposée dans le chapitre 2, en s'appuyant sur une étude clinique disponible publiquement [86] et en utilisant les technologies du Web Sémantique. Le modèle intègre des données multi-omiques analysées à l'échelle du patient. L'intérêt d'avoir des données multi-omiques est de pouvoir identifier des mécanismes complexes concernant un phénotype particulier.

Dans ce chapitre, la connexion entre les données de génomique et de transcriptomique sera exploitée à partir du modèle d'intégration. La première section va définir formellement la notion d'*expression Individually-Consistent Trait Loci* (eICTL), comme étant un lien potentiel de causalité entre la présence d'un SNP et la variation d'expression de son gène propre. Afin d'extraire ces informations des données intégrées dans le modèle transomique, la définition est ensuite transformée sous forme de requête SPARQL dans la section 3.1.3. La première partie de la section 3.2 présente l'application de la requête sur le modèle et l'identification des eICTLs candidats, caractérisés comme étant des éléments transomiques. La sous-section 3.2.2 montre l'évaluation de l'apport des eICTLs candidats par rapport aux données génomiques. Deux valeurs ajoutées des éléments transomiques sont mises en évidence : la réduction des dimensions des données et leur aspect plus discriminants dans la classification des patients. Enfin, la sous-section 3.2.3 présente la comparaison des eICTLs avec leur homologue statistique, les eQTLs.

3.1 expression Individually-Consistent Trait Loci (eICTLs) : Relation entre SNPs, variation d'expression et maladie

3.1.1 Influence de SNPs sur le phénotype

Depuis plusieurs années, de nombreuses études ont cherché à mettre en évidence le rôle des variations génétiques sur des phénotypes d'intérêt. Les régions génomiques ou loci possédant ces variations et statistiquement corrélées à un caractère d'intérêt complexe (facteur de risque, phénotype particulier...) sont appelés quantitative trait locus (QTLs). Les études d'association pangénomique (ou GWAS) sont des études à large échelle qui ont permis de cataloguer de nombreux QTLs associés à des caractères phénotypiques particuliers comme des maladies [88].

Cependant, l'identification de ces QTLs seule ne permet pas de comprendre le mécanisme moléculaire sous-jacent. La variation du niveau d'expression des gènes est aussi un élément moléculaire qui peut être interprété comme étant un caractère phénotypique complexe. Ainsi, les déterminants génétiques qui sont corrélés avec ces variations sont appelés eQTLs (expression Quantitative Trait Loci) [89]. Lorsque l'effet du variant intervient sur l'expression de son propre gène, il s'agit de *cis*-eQTLs mais s'il agit sur un gène distant, on parle de *trans*-eQTLs. Depuis 2010, le projet GTEx [1] (Genotype-Tissue Expression) est un catalogue contenant de nombreux eQTLs, mesurés dans 49 tissus non pathologiques et multi-ethniques [90]. Le fait de pouvoir filtrer les eQTLs avec les QTLs identifiés par les études GWAS suggère l'hypothèse que les variations d'expression de ces gènes jouent un rôle dans la maladie.

L'intérêt du calcul des eQTLs est de pouvoir combiner des données génomiques et transcriptomiques. En revanche, ils sont calculés sur une population par des analyses statistiques. Ainsi, la limite qui est liée à ce type d'analyse est le besoin d'un grand nombre d'échantillons, issus de populations relativement homogènes, afin d'obtenir une puissance statistique suffisante. Comme démontré dans le chapitre 1, les données à disposition dans cette thèse se caractérisent par des populations hétérogènes et un nombre d'échantillons petit. De plus, l'objectif de cette thèse est d'utiliser une approche centrée sur le patient afin de prendre en considération les caractéristiques de chaque patient.

3.1.2 Les SNPs associés à la variation d'expression à l'échelle du patient

Dans le chapitre précédent, le modèle final intégratif permet de relier des données appariées, multi-omiques, dont les valeurs de chaque entité ont été calculées à l'échelle du patient.

Du fait de l'appariement et du caractère multi-omique des données, il est possible d'exploiter ce modèle en recherchant des connexions potentielles entre la génomique et la transcripto-

1. https ://commonfund.nih.gov/GTEx/

mique. Dans cette optique, un nouveau type d'association est défini dans cette thèse, il s'agit des eICTLs. A la différence des eQTLs qui sont une association statistique basée sur un grand nombre d'échantillons, les eICTLs sont des associations calculées par raisonnement et centrées sur le patient. Il s'agit d'identifier les SNPs dont la présence va influencer son niveau d'expression chez certains patients. Tout comme les eQTLs, cette influence peut être locale (*cis*-) ou distante (*trans*-). Lors de cette thèse, seules les influences locales, directes ont été abordées.

La figure 3.1 est une représentation schématique des conditions pour identifier les eICTLs candidats. La première condition est qu'il existe un SNP (SNP_x) dont la présence sur son gène porteur ($Gene_y$) coïncide avec une diminution de l'expression de ce dernier par rapport à son expression dans la population de référence. La seconde condition est qu'il existe au moins un individu pour lequel l'absence de ce SNP sur le gène porteur ne coïncide pas avec une diminution de l'expression de ce gène (ou alors entraîne une sur-expression de celui-ci).

Afin de ne pas prendre en compte les eICTLs qui ne sont spécifiques qu'à un seul patient, il est nécessaire qu'au moins deux individus vérifient la première condition.

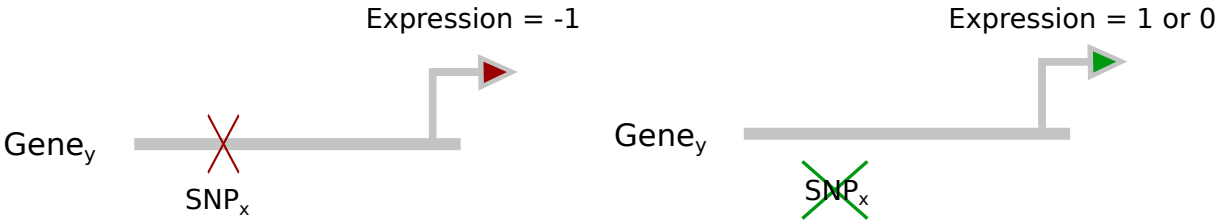


FIGURE 3.1 – **Représentation schématique des conditions pour l'identification des eICTLs candidats.** A gauche, la première condition est que la présence du SNP_x sur le gène $Gene_y$ entraîne la diminution de son expression par rapport à la population contrôle. A droite, la seconde condition où l'absence de ce SNP ne modifie pas l'expression du gène ou alors le sur-exprime.

De manière plus formelle, on appelle un eICTL candidat, un couple (SNP_x,$Gene_y$) pour lequel il existe une association directe telle que :

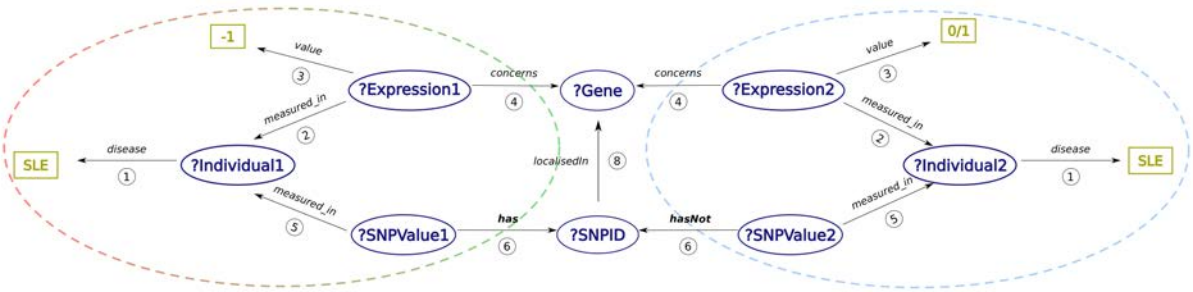
- $Gene_y$: gène dont l'expression a été mesurée et qui possède au moins 1 SNP
- SNP_x : SNP localisé sur un gène dont l'expression a été mesurée
- Il existe 3 patients, Indiv0, Indiv1, Indiv2 tels que :
  - Indiv0 vérifie : hasSnp(Indiv0, SNP_x) $\wedge$ Expr($Gene_y$,Indiv0) = -1
  - Indiv1 ($\neq$ Indiv0) vérifie : hasSnp(Indiv1, SNP_x) $\wedge$ Expr($Gene_y$,Indiv1) = -1
  - Indiv2 vérifie : ($\neg$ hasSnp(Indiv2, SNP_x)) $\wedge$ Expr($Gene_y$,Indiv2) $\in$ {0 ;1}

La recherche des eICTLs candidats dans un modèle comme celui présenté au chapitre 2 nécessite de transformer la définition précédente en requête SPARQL.

3.1.3 Définition des (eICTLs) candidats en langage SPARQL

Pour transformer la définition des eICTLs candidats en requête SPARQL, il est nécessaire dans un premier temps de faire le lien entre les entités et leurs relations (prédicats) dans le endpoint et les contraintes de la définition. Cela permet de ne sélectionner que les prédicats qui seront utiles dans la recherche de eICTLs candidats. Dans le cas présent, le nom des prédicats et des entités est assez similaire à ceux de la définition (ex : le SNP est localisé sur un gène = SNP *localisedIn* Gene) ce qui rend la sélection assez simple.

Ensuite, il est important de générer une requête performante d'un point de vue temps de calcul. Cet aspect a été plus compliqué à mettre en place. En effet, la structuration la plus intuitive est de réaliser la requête en un seul bloc. Cependant, comme le montreront les résultats dans la prochaine section lors de l'application, le temps de calcul est assez long. Il a donc été nécessaire de re-structurer la requête en deux sous-requêtes afin de diminuer le temps d'exécution.



(a) Représentation graphique de la requête

La requête se fait en deux parties. Tout d'abord, elle va lister les ensembles {gène, SNP, nombre d'individus} (ligne 12) pour lesquels les individus sont des patients (ligne 14) dont le gène a une valeur d'expression à -1 (ligne 15) et présente un SNP chez ces individus (ligne 16).

La ligne 31 permet d'éliminer tous les ensembles dont le nombre d'individus avec ces caractéristiques biologiques serait inférieur à 2.

Cette première partie va donc générer des résultats intermédiaires qui seront ensuite filtrés avec la deuxième partie de la requête. A partir de cette liste intermédiaire, la seconde partie de la requête ne va garder que des couples {gène, SNP} où le gène a une valeur d'expression à 0 ou 1 (ligne 37) et qui ne possède pas le SNP (ligne 38) chez d'autres individus qui sont eux aussi des patients (ligne 36).

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX ask: <http://askomics.org/data/>
SELECT DISTINCT ?Gene1_L ?SNP1_L
WHERE {

    ?Gene1 rdf:type ask:Gene .
    ?Gene1 ask:Type ask:protein_coding .
    ?Gene1 rdfs:label ?Gene1_L .
    ?SNP1 rdfs:label ?SNP1_L .

    # Sub-query 1 : Search for gene-SNP pairs associated to Indiv0 and Indiv 1 (red and green petal)
        {SELECT DISTINCT ?Gene1 ?SNP1  (COUNT(DISTINCT ?INDIV1) as ?I1) WHERE {
    VALUES ?INDIV1_DiseaseCategory { ask:SLE }
    VALUES ?Expression1_Disc { <http://www.semanticweb.org/user/ontologies/2018/1#-1> }

    ?Expression1 ask:Exp_concerns ?Gene1 .
    ?Expression1 rdf:type ask:Expression .
    ?Expression1 ask:Value ?Expression1_Disc .
    ?Expression1 ask:Exp_measuredIn ?INDIV1 .
    ?INDIV1 rdf:type ask:INDIV .
    ?INDIV1 ask:Disease ?INDIV1_DiseaseCategory .
    ?SNPValue1 ask:measuredIn ?INDIV1 .
    ?SNPValue1 rdf:type ask:SNPValue .
    ?SNPValue1 ask:has ?SNP1 .
    ?SNP1 ask:localisedIn ?Gene1 .
    ?SNP1 rdf:type ask:SNP .
    }}
    FILTER (( ?I1 > 2 ))

    # Sub-query 2 : Filtering gene-SNP pairs associated to Indiv2 (blue petal)
    FILTER EXISTS {
    VALUES ?INDIV2_DiseaseCategory { ask:SLE }
    VALUES ?Expression2_Disc { ask:1 ask:0 }
    VALUES ?Genotypage1_2Value { 0 1 }
    ?Expression2 ask:Exp_concerns ?Gene1 .
    ?Expression2 rdf:type ask:Expression .
    ?Expression2 ask:Value ?Expression2_Disc .
    ?Expression2 ask:Exp_measuredIn ?INDIV2 .
    ?INDIV2 rdf:type ask:INDIV .
    ?INDIV2 ask:Disease ?INDIV2_DiseaseCategory .
    ?SNPValue2 ask:measuredIn ?INDIV2 .
    ?SNPValue2 rdf:type ask:SNPValue .
    ?SNPValue2 ask:hasNot ?SNP1 .
    ?SNP1 ask:localisedIn ?Gene1 .
    ?SNP1 rdf:type ask:SNP .}}
```

(b) Requête en langage SPARQL

FIGURE 3.2 – **Différentes représentations de la requête SPARQL permettant d'identifier les eICTLs candidats chez les individus SLE tels que les gènes ont au moins une expression à -1 et possèdent 1 SNP.** (a) Chaque ellipse est associée à une sous-requête. L'ellipse de gauche correspond à la sous-requête 1 qui va identifier un premier ensemble de couple gène-SNP dont le gène est sous-exprimé chez au moins 2 patients SLE et qui est associé à un SNP (rouge et verte). L'ellipse de droite correspond à sous-requête 2 qui va filtrer ce premier ensemble en ne gardant que les couples gène-SNP dont le gène n'est pas sous-exprimé et qui ne possède pas le SNP (ellipse bleue). (b) Requête SPARQL de la recherche de eICTLs candidats

3.2 Application de requête sur le modèle de données intégrées transomique

3.2.1 Identification des eICTLs candidats

Comme expliqué dans la section précédente, la première requête SPARQL générée était réalisée en un seul bloc. Cela sous-entend que la requête recherche l'ensemble des couples gène-SNP existant chez les patients SLE, en prenant les valeurs fixes définies à la figure 3.2. Puis, à la fin, elle réalise le filtre sur le nombre d'individu ayant l'expression à -1 et contenant le SNP. Le temps de calcul était de près d'1h (Figure 3.1).

	Structure de la requête	Temps de calcul	Nombre de eICTLs candidats
Requête 1	un seul bloc	58 min	1,961
Requête 2	2 sous-requêtes	30 sec	1,961

TABLE 3.1 – **Comparaison des performances selon les différentes structurations de la requête.** La requête 1 correspond à la première structuration de la requête en un seul bloc. La requête 2 est cette même requête après étape d'optimisation.

Une étape d'optimisation de la requête a donc été nécessaire. Une première hypothèse afin d'obtenir de meilleures performances est de découper la requête initiale en deux sous-requêtes. La première sous-requête réalise un premier filtre sur les couples gène-SNP afin de ne garder que ceux dont l'expression du gène est -1 sachant qu'il présente un SNP. Cette sous-division de la requête permet de réduire le temps d'exécution par 100 (Figure 3.1).

Dans les deux cas, la requête a permis l'identification, sur la totalité du endpoint, de 1,961 eICTLs candidats. Ainsi, sur l'ensemble des 6,106 SNPs, il n'en existe que 1,961 qui peuvent avoir une influence sur la diminution d'expression du gène chez au moins deux individus SLE. Cette requête permet de sélectionner un ensemble de SNPs qui semblent plus informatifs.

Ces résultats peuvent être utiles afin de réduire la dimension des données génomiques. Après leur identification, il est nécessaire d'évaluer ces eICTLs candidats en tant que marqueurs potentiels pour les patients SLE.

3.2.2 Comparaison des analyses génomiques et transomiques

Les liens entre génomique, transcriptomique et transomique - Les données d'expression de gènes permettent d'obtenir un signal fort qui est associé à la pathologie. Plus précisément, ce signal peut s'interpréter comme un mélange entre les causes propres de la maladie (dérégulations initiales) et les conséquences qui en découlent. Les données génomiques, quant à elles, peuvent être utilisées pour rechercher exclusivement la cause de la maladie. Cependant, le signal peut être beaucoup plus faible et diffus.

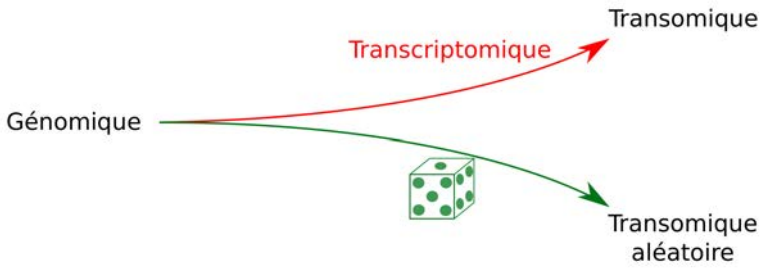


FIGURE 3.3 – **Représentation du filtrage des données de génomique par la transcriptomique selon la définition des eICTLs candidats ou par un filtrage aléatoire.** Les eICTLs candidats sont des éléments transomiques et définis comme étant des SNPs qui ont une influence sur l'expression de leur gène. Il s'agit donc d'un sous-ensemble de données de génomique, filtrée par les données de transcriptomique.

La recherche des eICTLs candidats permet de ne garder qu'un sous-ensemble des données de génomique en les filtrant grâce aux données de transcriptomique en suivant la définition formelle décrite précédemment.

L'évaluation de l'apport de la transomique a donc été réalisée en comparant le filtrage de la génomique par la transcriptomique par rapport à un filtrage aléatoire. Les résultats de ce dernier sont appelés transomique aléatoire. (Figure 3.3).

	SNP1	SNP2
I1	C	C
I2	NC	NC

Genomic
147 indv × 6 197 variables

	Gene1	Gene2
I1	Down	Normal
I2	Down	Normal

Transcriptomic
147 indv × 40 834 variables

	SNP1-Gene1	SNP2-Gene2
I1	VP	FN
I2	FP	VN

Transomic
147 indv × 1 961 variables

FIGURE 3.4 – **Exemple des trois types de matrices (avec leur tailles) utilisées pour comparer les données de génomique, transomique et transcriptomique.** Pour les trois matrices, les lignes représentent les individus et les colonnes représentent les variables. En génomique, les variables sont les SNPs et les modalités sont *Contains - C* ou *NotContains - NC*. En transcriptomique, les variables sont les gènes et les modalités peuvent être {*Down, Normal, Up*}. En transomique, les variables sont les eICTLs et les modalités peuvent être {*VP, VN, FP, FN*}.

Réduction des dimensions des données - Les résultats transomiques sont présentés sous forme de matrice catégorielle à 4 modalités, celle de génomique à 2 modalités et celle de trans-

criptomique à 3 modalités. Pour les 3 matrices, les individus sont en ligne et les variables explicatives en colonne. La figure 3.4 montre un exemple pour chaque matrice. Les modalités en génomiques retranscrivent la relation *has* (*Contains - C*) ou *hasNot* (*NotContains - NC*) de la requête SPARQL. En transcriptomique, les valeurs discrétisées sont représentés par les 3 valeurs {*Down, Normal, Up*}. Concernant les résultats de transomique, les modalités sont 4 règles définies comme suit :

- *VP* : La présence du SNP est associée à une sous-expression de son gène
- *VN* : L'absence du SNP est associée à une expression normale ou sur-expression de son gène
- *FP* : L'absence du SNP est associée à une sous-expression de son gène
- *FN* : La présence du SNP est associée à une expression normale ou sur-expression de son gène

Comme le montre la taille des différentes matrices, la sélection des eICTLs permet de réduire le nombre de variables explicatives en ne gardant que les SNPs ayant un potentiel informatif sur l'expression de leur gène. Pour les résultats de transomique, le nombre des variables explicatives est réduit par 3 par rapport aux données de génomique.

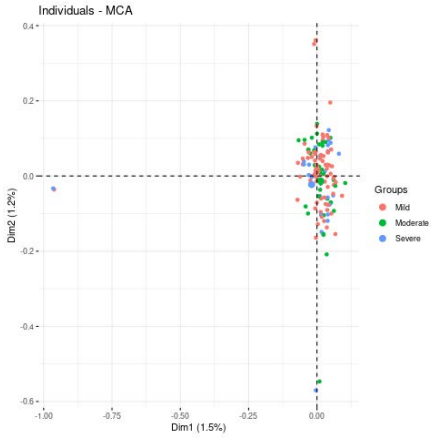
Dans la suite de l'analyse, ce potentiel est évalué en comparant les résultats de transomique, génomique et transcriptomique.

Analyse des Correspondances Multiples (MCA) - L'Analyse des Correspondances Multiples est un dérivé de l'analyse en composante principale (ou PCA) qui s'applique aux données qui ne sont pas numériques mais catégorielles. Les deux premières dimensions affichées sur les graphes représentent les deux premiers ensembles de variables expliquant le maximum d'informations sur la population. Plus le pourcentage d'information est grand, plus les données analysées sont séparées. Le package *FactoMineR* [91] en R a été utilisé pour réaliser les analyses de MCA.

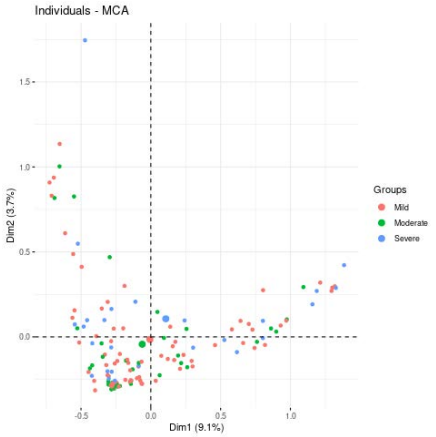
L'analyse des données de génomique montre que les deux premières dimensions ne permettent d'expliquer que 3% de l'information des patients SLE (Figure 3.5a). On peut voir que la majorité de ces patients restent regroupés au centre des deux axes.

Les individus sont plus dispersés lorsque la MCA est appliquée sur les résultats transomiques issus de la requête (Figure 3.5b). On constate en effet que les deux dimensions sont plus informatives qu'en génomique seule (plus de 12%). Cette différence ne s'explique pas uniquement par la réduction de dimension car, en comparaison, une sélection aléatoire des eICTLs candidats ne permet d'expliquer qu'un peu plus de 9% de l'information. (Figure 3.5c).

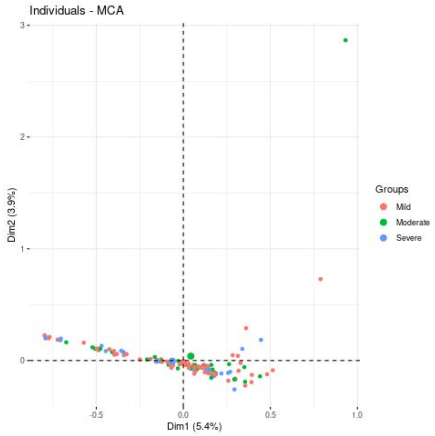
On note que la transcriptomique est la couche qui semble la plus informative concernant les patients. En effet, la première dimension explique près de 60% des données (Figure 3.5d). Cela s'explique par le fait que les DEGs des données d'expression sont un signal fort dans la population.



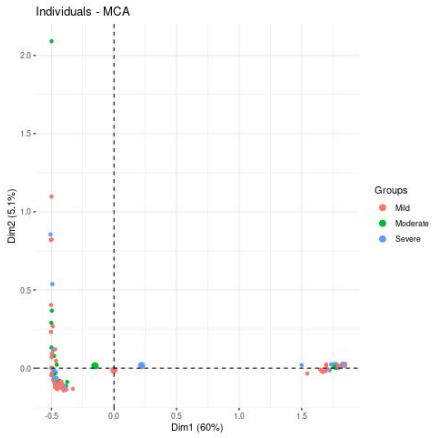
(a) MCA sur les données de génomique



(b) MCA sur les données de transomique



(c) MCA sur les données de transomique aléatoire



(d) MCA sur les données de transcriptomique

FIGURE 3.5 – **L'Analyse des Correspondances Multiples (MCA) sur les matrices des données de génomique, transomique, transomique aléatoire et transcriptomique, uniquement pour les patients SLE.** Chaque figure correspond aux résultats de MCA sur les différentes couches omiques. Sur le même principe que l'analyse en composante principale (PCA), les axes représentent les deux dimensions qui expliquent le plus les données et en quel pourcentage. Chaque point représente un individu SLE et la couleur correspond à la catégorie SLEDAI pour laquelle il est associé (rouge pour mild, vert pour moderate et bleu pour severe).

Clustering hierarchique des résultats de MCA - La fonction *HCPC* du package *FactoMineR* permet de réaliser un clustering hiérarchique sur les composantes principales issues de la MCA. De plus, l'option *nb.clust = -1* découpe de manière automatique l'arbre selon le nombre de clusters qu'il considère comme optimal. C'est l'option qui a été choisie pour le reste de l'analyse.

Cinq clusters sont identifiés en utilisant les données de génomique (Figure 3.6a) alors qu'il n'y en a que trois avec les données transomiques. Cependant ces derniers clusters semblent plus robustes que ceux de génomique puisque la distance des branches ou distance cophenétique est plus grande en transomique (Figure 3.6b).

En comparaison, les six clusters issus de la transomique aléatoire semblent moins robuste que ceux de transomique au vu des distances entre les clusters (Figure 3.6c).

Le clustering des données issues de l'analyse transcriptomique montre 5 clusters assez robuste en regardant les valeurs des branches (Figure 3.6d).

Les résultats transomiques permettent donc d'identifier moins de clusters qu'en génomique mais avec une meilleure définition de ces clusters.

Comparaison du clustering des patients - L'ensemble de ces clusters représente une nouvelle séparation des patients selon leur similarité pour chaque couche -omique.

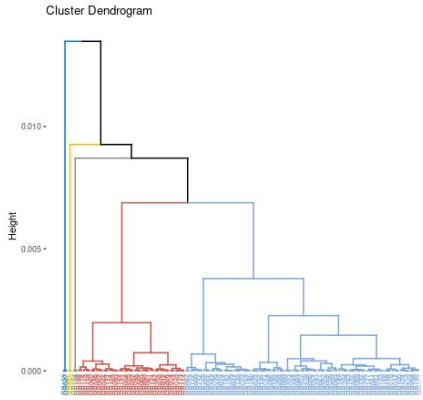
Le diagramme de Sankey permet d'analyser, dans ce contexte, la proportion de patients qui apparaissent dans des clusters différents selon les données -omiques (Figure 3.7). Chaque cluster est ordonné d'après les résultats du clustering hiérarchique précédent.

Le plus grand cluster de transomique permet de regrouper plus de la moitié des patients des deux plus grands clusters de génomique ainsi que les plus petits (Figure 3.7a). Le reste des deux grands clusters de génomique se rassemblent d'abord dans le cluster 3 puis 1 de transomique. Ce diagramme montre aussi que les clusterings transomique et transcriptomique sont assez similaire sur leur gros groupes. Les deux groupes transcriptomiques supplémentaires sont intégrés dans le cluster 2 transomique. Cela s'explique par le fait que les données transomique sont influencées par celles de transcriptomique, qui joue le rôle de filtre sur les données de génomique.

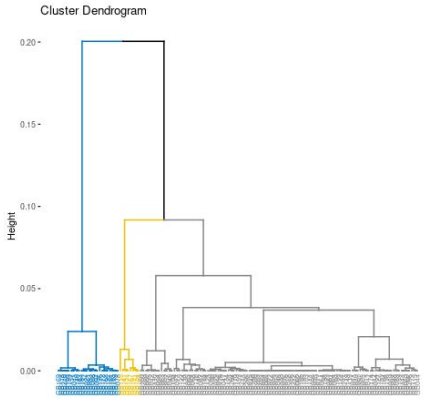
Si le résultat de transomique aléatoire est assez similaire au résultat de transomique, on note l'apparition de quelques très petits clusters. (Figure 3.7b).

Afin de poursuivre cette étude, nous avons cherché à évaluer la robustesse des groupes de patients mis en évidence *via* un ré-échantillonnage du jeu de données.

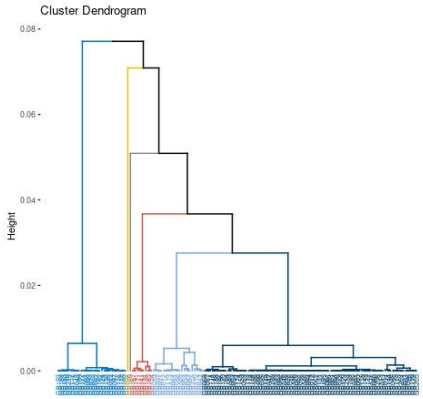
Robustesse du clustering - Afin de vérifier la robustesse des clusters dans les différentes couches -omiques, les analyses de MCA suivies du clustering hiérarchique ont été répétées sur plusieurs ré-échantillonnages de la population : 10 jeux de données avec 100, 120, 140, 160 et 180 individus chacun. Du fait de leur faible nombre, les individus sains ne sont pas ré-échantillonnés (tous les individus sains et ré-échantillonnage parmi les patients).



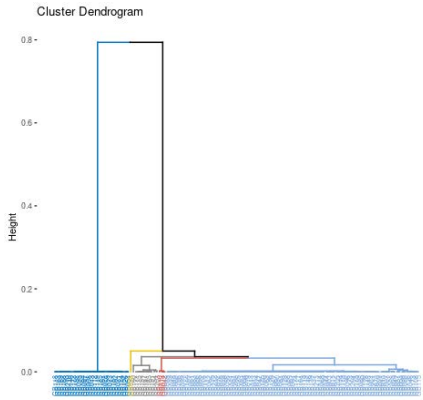
(a) Clustering hiérarchique des données de génomique



(b) Clustering hiérarchique des données de transomique

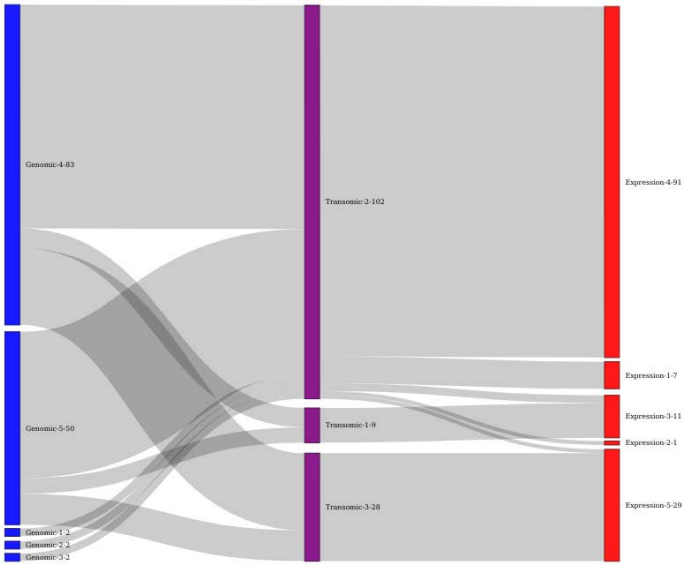


(c) Clustering hiérarchique des données de transomique aléatoire

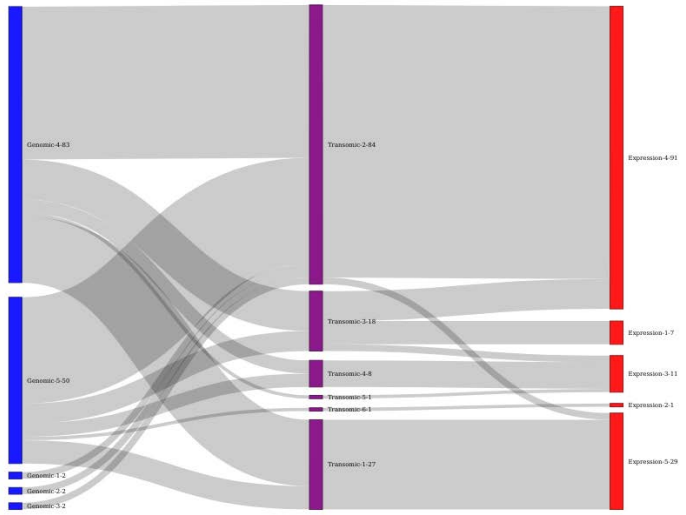


(d) Clustering hiérarchique des données de transcriptomique

FIGURE 3.6 – **Clustering hiérarchique des patients SLE selon les résultats de la MCA pour les couches de génomique, transomique, transomique aléatoire et transcriptomique.** Chaque clustering est réalisé à la suite des analyses de MCA décrite à la figure 3.5, par la fonction *HCPC* du package *FactoMineR*. Le découpage des clusters est réalisé de manière automatique par la fonction *HCPC*.



(a) Comparaison des clustering de patients entre les données : génomique - transomique - transcriptomique



(b) Comparaison des clustering de patients entre les données : génomique - transomique aléatoire - transcriptomique

FIGURE 3.7 – **Diagramme de Sankey représentant l'évolution du clustering des patients selon les différentes données omiques.** Chaque colonne correspond aux différentes couches omiques (génomique en bleu, transomique en violet et transcriptomique en rouge). Chaque bloc en colonne représente un cluster. L'organisation des clusters est celle définie par les résultats du clustering hierarchique en figure 3.6. La taille de ces blocs est proportionnelle aux nombres de patients contenus dans les clusters. Les flux entre les colonnes représentent les proportions de patients qui se retrouvent dans les clusters de la couche omique suivante.

A partir des résultats de clustering de chaque couche -omique ré-échantillonnée, nous avons généré une matrice de co-clustering. Au sein de cette matrice, chaque élément $M_{(i,j)}$ mesure le nombre de fois où la paire de patients SLE (i, j) apparaît dans un même cluster, normalisé par le nombre de fois où i et j faisaient partie des échantillons crées par ré-échantillonnage. Une heatmap a été créée à partir de chacune de ces matrices (Figure 3.8).

La matrice de co-clustering calculée grâce aux données génomique montre que les 3 plus petits clusters semblent être les plus robustes (Figure 3.8a). En revanche, même s'il est possible de distinguer les deux carrés représentant les deux plus grands clusters, la robustesse de ces clusters n'est pas aussi franche que pour les trois autres groupes.

Dans la matrice de co-clustering calculée grâce aux données transomiques, les 3 clusters semblent plus homogènes que ceux issus des résultats de génomique (Figure 3.8b). Les deux clusters aux extrémités de la heatmap (rouge et bleu) sont les plus homogènes des trois. On retrouve dans le plus grand cluster certains patients qui ne sont pas systématiquement associés au même cluster mais de façon plus atténuée que pour les résultats de génomique. En revanche, lorsque le choix des eICTLs candidats est effectué de manière aléatoire, les clusters ne semblent pas résister au ré-échantillonnage (Figure 3.8c).

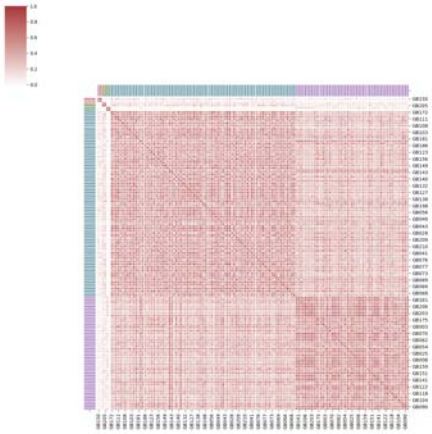
L'appartenance de deux patients SLE aux mêmes clusters lors du ré-échantillonnage semble plus robuste en transcriptomique qu'en génomique (Figure 3.8d). En effet, les rectangles des différents clusters sont mieux définis que précédemment. Mis à part dans le cluster violet qui est homogène, les autres clusters ne semblent cependant pas aussi homogène. En effet, certaines lignes dans les autres clusters apparaissent plus claires. Cela signifie que les couples de patients correspondants apparaissent moins souvent associés dans le même clusters. Cet aspect apparaît le plus dans le cluster rouge, en haut à gauche.

En conclusion, les résultats de génomique seule ne permettent pas d'obtenir des clusters suffisamment robustes. En comparaison, notre définition d'eICTLs nous permet de diminuer la dimension en générant des clusters assez robustes lors d'une analyse par ré-échantillonnage du jeu de données. Cette robustesse n'est pas simplement due à la réduction de dimension puisque les clusters générés par transomique aléatoire (sélection d'un sous-ensemble aléatoire du même nombre d'élement qu'en transomique) sont nettement moins robustes.

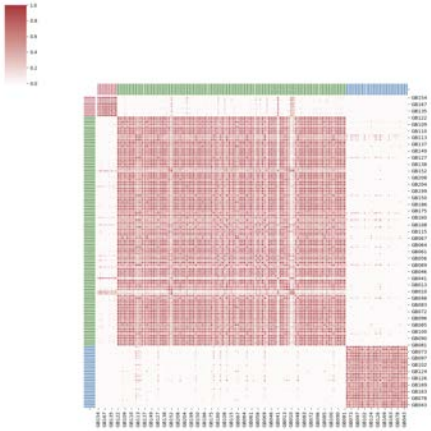
3.2.3 Comparaison avec les eQTLs (GTEx)

Comme défini dans les deux premières sections, les eICTLs candidats sont des associations SNP-variation d'expression qui sont proches des eQTLs dans leur définition. Nous proposons donc de comparer les eICTLs avec les eQTLs déjà identifiés.

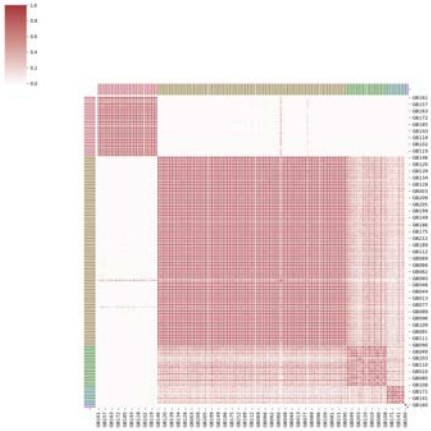
De nombreux eQTLs sont stockés dans la base de données GTEx qui est structurée par tissu. Dans leur publication de 2015, les auteurs ont justifié cette organisation car les maladies complexes ont souvent une origine dans la dysfonction de plusieurs tissus ou lignées cellulaires [92].



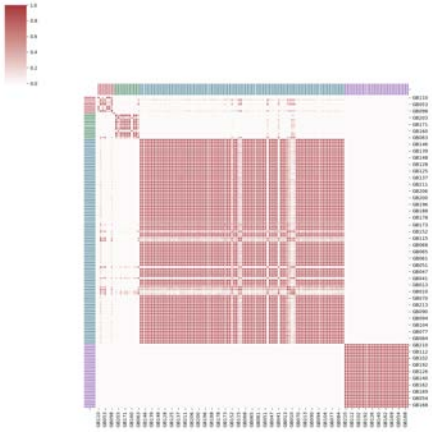
(a) Genomique



(b) Transomique



(c) Transomique aléatoire



(d) Transcriptomique

FIGURE 3.8 – **Les matrices d'appartenance d'une paire de patients au même cluster en génomique, transomique et transcriptomique.** Chaque ligne correspond a un patient SLE. Ils sont triés selon leur appartenance aux différents clusters. La même organisation est appliquée aux mêmes patients en colonne. Chaque valeur de la heatmap correspond au ratio pour lequel les deux patients sont associés aux même clusters sachant qu'ils apparaissent ensemble dans le dataset rééchantilloné. S'ils sont toujours associés au même cluster à chaque fois qu'ils apparaissent dans le même dataset, alors la valeur du ratio est de 1.

Des études combinent les QTLs mesurés dans les analyses GWAS pour une pathologie donnée et les eQTLs de cette base de données. Cela permet de valider d'une part les SNPs identifiés comme ayant une influence potentielle. D'autres part, les relations qui découlent des SNPs issus des GWAS et leur implication dans la variation d'expression d'un tissu avec les eQTLs donnent des informations sur certains mécanismes de régulation de la pathologie [93], [94].

Dans le cas des eICTLs candidats, les relations SNPs-variation d'expression-pathologie sont déjà mesurés en une seule requête. La grande différence entre les eICTLs candidats et les eQTLs est la mesure à l'échelle du patient des eICTLs candidats. Il apparaît donc important de vérifier si les eICTLs candidats ont déjà été caractérisés comme étant des eQTLs. Les données de la thèse étant réalisées dans le sang complet, seul ce tissu a été pris en compte dans la comparaison.

La figure 3.9 montre la proportion des eICTLs candidats associée ou non avec des eQTLs référencés au sein de la base de données GTEx.

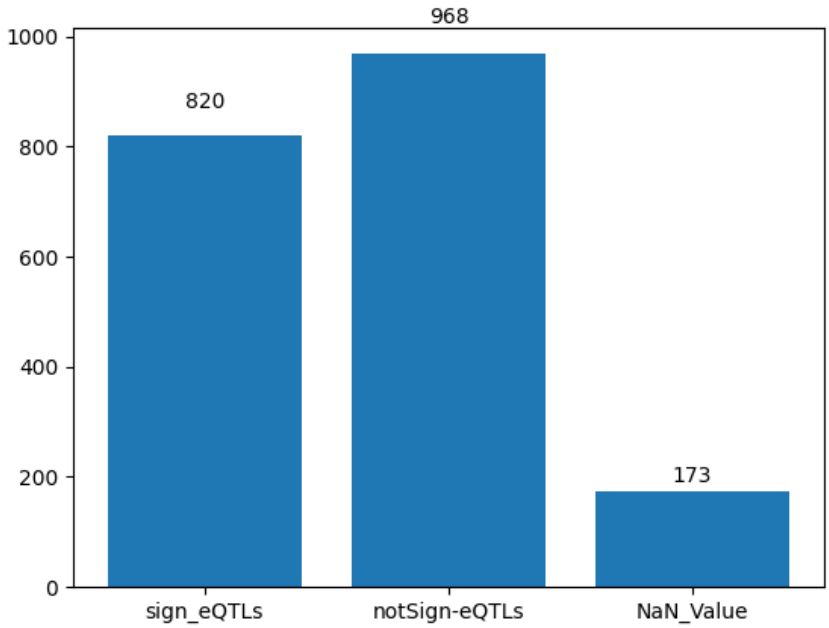


FIGURE 3.9 – **Proportion des eICTLs candidats identifiés comme eQTLs sur le portail GTEx (v8) pour le sang complet ("*Whole Blood*")**La première colonne représente le nombre de eICTLs candidats qui ont été identifiés comme des eQTLs testés dans le sang et qui sont significatifs. La seconde colonne correspond au nombre de eICTLs candidats qui ne sont pas retrouvés en tant que eQTLs significatifs. La troisième colonne est le nombre de eICTLs candidats qui n'ont pas été mesurés en tant que eQTLs dans le sang ou dont le variant et/ou le gène ne sont pas reconnus par GTEx.

Sur la totalité des eICTLs candidats identifiés, 9% correspondent à des eICTLs candidats qui n'ont pas pu être retrouvés dans la base de données (gène ou SNP sans correspondance) ou qui n'ont pas encore été mesurés dans le sang.

Parmi les 1 788 eICTLs candidats retrouvés, près de 46% sont associés à des eQTLs significatifs (p-value <5%). Cette p-value est mesurée pour chaque paire variant-gène de manière à tester l'hypothèse alternative selon laquelle le coefficient de corrélation entre le variant et la variation d'expression du gène est différente de 0 et ce dans chaque tissu [92].

Cela signifie que 54% des eICTLs candidats restants sont, dans GTEx, des eQTLs considérés comme non significatifs. Il pourrait donc s'agir de couples qui seraient spécifiques aux patients atteints de SLE.

En conclusion, plusieurs eICTLs candidats (46%) ont été retrouvés comme des eQTLs significatifs, mesurés dans le sang, tandis que d'autres (54%) ont pu être associés à des eQTLs non significatifs dans le sang. De plus, certains liens n'ont pas encore été mesurés dans ce même tissu. L'hypothèse au vu de ces résultats est que les couples gène-SNP qui ne sont pas des eQTLs significatifs seraient spécifiques à certains patients SLE.

3.3 Conclusion : eICTLs candidats comme nouveaux marqueurs de patients

L'objectif du chapitre 3 était de pouvoir exploiter le modèle défini au chapitre précédent afin d'identifier des éléments pouvant expliquer les mécanismes moléculaires causaux de la pathologie tout en prenant en compte les caractéristiques des patients.

Un eICTL candidat est défini formellement comme étant un couple SNP-gène dont la présence du SNP a une influence directe potentielle sur la variation d'expression de ce gène. L'interrogation du modèle transomique par l'intermédiaire d'une requête SPARQL complexe a permis l'identification de 1 961 eICTLs candidats. Une étape d'optimisation de la requête a permis de réduire le temps d'exécution par 100. La première valeur ajoutée des eICTLs candidats est la réduction de dimension des données puisque cette sélection de SNP permet de réduire par 3 fois la quantité de données de génomique. La seconde valeur ajoutée est leur apport informatif car l'évaluation de ces liens a montré qu'ils sont plus discriminants dans la caractérisation des échantillons que les données de génomique ou une selection aléatoire, puisqu'ils génèrent des clusters plus robustes.

Les eQTLs sont mesurés par association statistique entre la présence d'un SNP et la variation d'expression de son gène. Les eICTLs candidats sont eux calculés par raisonnement et prennent en compte l'hétérogénéité de la population. La comparaison entre les deux a montré que la majorité des eICTLs candidats sont retrouvés en tant qu'eQTLs. Parmi eux, la moitié sont significatifs dans le sang complet (couple variant-gène souvent corrélés dans ce tissu) et l'autre moitié ne l'est pas. Ces eICTLs candidats pourraient être des couples SNP-gène spécifiques de sous-groupes de patients SLE.

Ainsi, ces éléments transomiques pourraient être de nouveaux biomarqueurs caractéristiques

de sous-populations de patients SLE.

Les effets causaux peuvent aussi être inférés par l'utilisation de réseaux de régulation de gènes. Cette connaissance supplémentaire permettrait, à partir d'une liste de marqueurs identifiés par des données d'expression par exemple, de remonter l'information de causalité en prenant en compte les dépendances de régulation entre les gènes. Cette approche sera expliqué dans le chapitre suivant.

Chapitre 4

Signature complexe par analyse des connaissances

La recherche de causalité ne se limite pas à l'utilisation de données -omiques, il est aussi important de prendre en compte les liens entre les entités.

Ce chapitre présente une méthode permettant de classer automatiquement des états stables d'un réseau booléen en utilisation l'analyse par concept formel (FCA).

4.1 Présentation de l'article

L'article présenté dans ce chapitre a fait l'objet d'une publication dans le *Journal of Theoretical Biology* en 2019 (doi :https://doi.org/10.1016/j.jtbi.2019.01.015).

La biologie des systèmes est un domaine qui permet de mieux comprendre le comportement que peut générer un réseau biologique. Ce comportement peut être représenté sous forme d'états stables du système.

Selon la taille et la complexité de ces systèmes, le nombre d'états stables générés peut être très importants. Il apparaît nécessaire de pouvoir interpréter ces états fixes de manière automatique. L'approche proposée est de classer ces états en fonction de certains noeuds actifs. Ces activations sont associées à des signatures préalablement définies par des expertises biologiques et caractéristiques de phénotypes biologiques connus.

C'est dans cet optique qu'a été développée la méthode présentée dans ce chapitre. Elle a été appliquée sur un réseaux de régulation de gènes caractérisant la différenciation des lymphocytes T. Les différents comportements du système sont interprétés comme étant une différenciation en plusieurs types lymphocytaires. Elle s'articule autour de 3 axes importants :

— la classification automatique des états stables ;
— l'enrichissement de signatures associées à des phénotypes ;
— l'identification de nouvelles signatures hybrides

Pour réaliser ces objectifs, la méthode s'appuie sur une approche de bi-classification symbolique : l'analyse des concepts formels (*Formal Concept Analysis* ou FCA) qui permet de

représenter les associations au sein d'une matrice au sein d'une structure hiérarchique composée de bi-clusters appelés "concepts". Chaque concept est constitué d'un ensemble de lignes de la matrice (dans notre cas, il s'agit d'états stables) qui partagent des caractéristiques similaire vis-à-vis d'un certain nombre de colonne (dans notre cas, il s'agit des composés biologiques activés dans les états stables considérés). L'intérêt de la FCA pour classer les états stables est d'obtenir une liste exhaustive des clusters. En effet, les approches classiques de classification ne fourniraient qu'un sous-ensemble de ces clusters. Une comparaison a été réalisée dans ce sens entre une approche classique, le clustering hiérarchique et la FCA, en basant la classification sur deux signatures de tailles différentes. Les résultats montrent que le clustering hiérachique va perdre l'information de la seconde signature qui est noyée dans la première, plus grande. Cette perte s'explique par le fait que certains états stables contiennent des éléments des deux signatures et vont préférentiellement être associés au signal le plus fort.

L'automatisation de la génération et de la classification des états stables va mettre en avant deux points importants.

Le premier concerne l'ensemble des états stables d'un cluster associé à une signature et dont certains des noeuds s'activent systématiquement en même temps que les éléments biologiques constituant cette signature. Les signatures canoniques définies au départ sont donc enrichies grâce aux dépendances de régulation du système. De plus, certains sous-regroupements d'états stables partagent des formes dérivées des signatures canoniques et sont appelés variants. Parmi ces variants, ceux associés à deux signatures canoniques sont appelés signatures hybrides.

Le second permet de prédire de nouvelles signatures composées de deux signatures canoniques, les signatures hybrides. La génération et la classification automatique d'un ensemble d'états stables issu d'un réseau de régulation de gènes a permis l'identification d'une nouvelle signature. En effet, en comparaison avec une étude dont la classification s'est faite manuellement, la méthode a retrouvé l'ensemble des groupes automatiquement et a identifié une classe d'états stables associée à une nouvelle signature hybride.

4.2 Introduction

Systems biology aims to understand how the interactions between cellular components determine the cell response to environmental perturbation by external stimuli. Historically, two main approaches have been developed to take into account the dynamical behavior of a regulatory network. Inspired by modeling technics used in physics, continuous models based on differential equations are widely used to investigate the role of circuits in regulatory and signaling networks, as well as the fluctuations of concentrations over time. Notice however that model calibrations are difficult and require a large set of quantitative experimental measurements which are hardly available in the context of regulatory interactions [20], [21], [95]. Another weakness of continuous and stochastic modeling technics is that they implicitly assume that reactions follow mass action kinetics, although regulatory interactions have been observed and measured to be rather similar to switches, or at least very sharp in terms of responses [96], [97].

To overcome these limitations, discrete frameworks have been introduced to describe the response of regulatory networks. A first class of synchronous models associate the gene or protein states to binary variables whose values are controlled by the binary values of their regulators [22]. This approximation of regulation appeared to be overly simplistic because it does not take into account the possibly different time-scales occurring in regulatory networks, nor the fact that a component might act differently according to its level of expression or activation. Multi-level and asynchronous logical formalisms, the so-called Thomas models [23], [24], have been proved to be accurate for modeling regulatory systems because they capture the main features of the different time-scales in regulatory processes with asynchronous and non-deterministic formalisms [21], [98].

An output of the study of a logical network by multi-level and asynchronous logical formalisms is the enumeration of its steady states and more generally the study of its attractors [25]. This feature has been widely used to link models of regulations with phenotypes, especially in health-related applications. In [99], it was proved that the attractors of a mammal cell-cycle in several perturbed conditions were in agreement with known phenotypes in the literature. The steady states and attractors of a logical model were also proved to fit with genotyping information in [100]. Finally, in [101], [102], the authors studied a network of T-helper lymphocytes and evidenced that the steady states of the network in several environmental or gene-deletion/activation conditions were in agreement with observed clinical phenotypes. These phenotypes were either generic (proliferation, apoptosis...) or were more specific and described subtle differences in cancer cell-types.

All these studies establish a link between some phenotypes (especially in cancer situations) and the steady states of a logical network. In concrete terms, this link provides a signature for each phenotype, that is, a set of biological markers present in the steady states whose activation is characteristic of the phenotype.

Notice however that the concept of signature is loosely defined in the literature depending on the context. A phenotype signature is generally defined as the set of master genes or proteins characterizing this phenotype. Signatures can be computed according to gene set enrichment analysis and gathered in databases such as SigDB [103]. They can also be computed to focus on causality effects if logical networks are available [104]-[106]. However, most of the time, signatures are refined manually by clinicians or biological experts in order to be more accurate and discriminate cell-types according to a few biomarkers. This is for example the case of CD4 T helper cells (Th cells) for which several cell-types have been identified (Th1, Th2, Th17, Th9, Th22). The heterogeneity of Th cells is closely related to signals from the microenvironment ; typically IL-12 is required for the development of Th1 cells ; IL-4/IL-2 drive the development of Th2 subtype, and TGFb induces the differentiation of Th17 cells as well as T regulatory (Treg) cells. Moreover those main Th cells subtypes are associated with very specific biomarkers. Indeed Th1, Th2, Th17 are characterized by the expression of T-bet, Gata3 and RORgt respectively ; while Treg cells are characterized by Foxp3 expression. Yet, Th cells present a certain degree of plasticity, and notably they can adopt hybrid phenotypes.

Hybrid phenotypes can appear when the biomarkers of several signatures are measured simultaneously in the same cell-type [107]. For instance the Th1-Th17 subtype expresses both T-bet and RORgt master regulators and produces both INFg and IL-17. Therefore, links between signatures, phenotypes and steady states in logical models become intricate as soon as the network size increases. In this situation, classification methods such as hierarchical clustering [108] highlight the main links between signatures and phenotypes, but fail to describe all the possible variants and hybrids that co-exist with the main clusters.

Our study aims at developing an automatic method to classify the steady states of a logical network according to a given family of phenotypes. These phenotypes are defined by their own signatures that can be either a single master gene or more generally a pattern of activated biological compounds (genes, proteins or markers). Our framework allows a systematic exploration of the links between phenotypes signatures and values of genes in the steady states of a Boolean network.

It relies on a classification of the steady states using a hierarchical structure derived from Formal Concept Analysis (FCA), a data analysis method handling binary matrices [109]. In our case, we focus on the analysis of matrices describing the list of activated compounds in steady states. The FCA method produces a lattice, representing in a hierarchical way associations as bi-clusters of specific nature, named *concepts*. Each concept is made of a subset of rows (in our cases, steady-states) which exhibit similar characteristics across a subset of columns (in our case, the compounds shared by the steady-states). By studying concepts associated with signatures of phenotypes, thanks to this hierarchical structure, we can perform a systematic characterization of biological compounds which are always paired with the signature's master

genes or proteins according to the Boolean network. Hybrid phenotypes can also be characterized in association with all their possible variants. This is illustrated on several models representing the differentiation of LTh cells [101], [102], [110]. Interestingly, the classification enables the identification of novel hybrid types together with the simulation conditions that generated them.

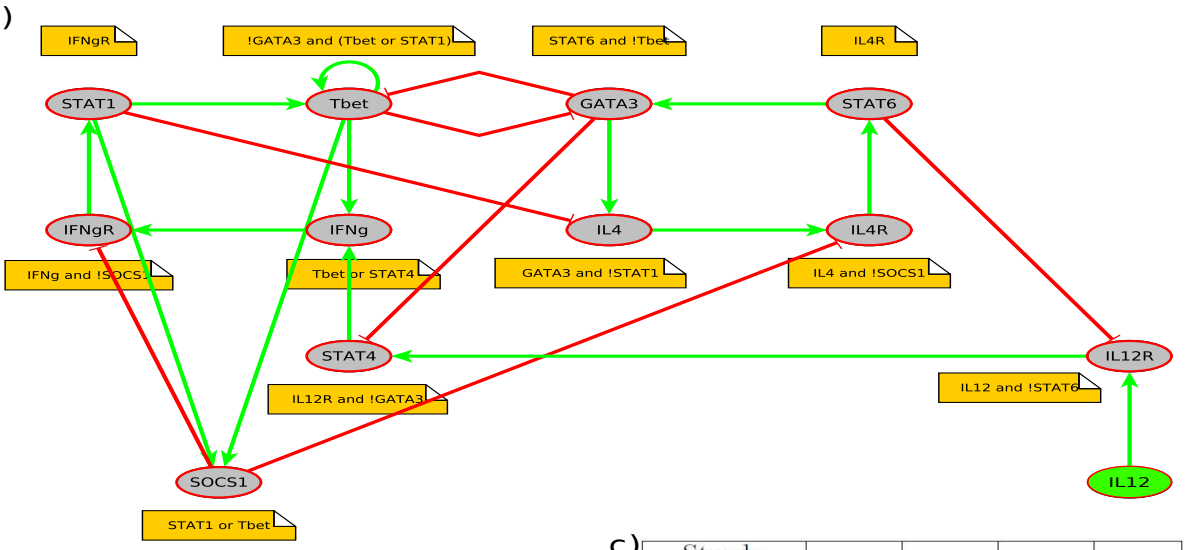
4.3 Organizing the steady states of a Boolean network into a lattice

4.3.1 Network representation and simulation

R. Thomas has proposed a logical formalism [23], [24] to model regulatory networks. It is based on two directed graphs and a system of logical rules coding for the network dynamics. The interaction network is represented by a *regulatory graph* (RG), whose nodes stand for the biological compounds of the system, and edges stand for the interactions between these components (transcriptional activation or inhibition). In this work, we distinguish two types of nodes, external nodes (also called input nodes) and internal nodes. External nodes represent the *input* of the conditions of simulation. These compounds can not be regulated during the dynamics but their values can be fixed. To each internal compound are attached (1) a discrete variable representing the expression of the biological compound qualitatively (its *state*). We consider here only Boolean variables ; (2) a logical function depicting the evolution of the component with respect to the states of its regulators. If the values of the internal compounds are specified at the beginning of the simulation, this state is called an *initial state*. The *State Transition Graph* (STG) represents the discrete dynamics ; nodes are the states of the system, and transitions link two consecutive states [23]. Hence, the STG encompasses all the possible trajectories with respect to the set of logical functions parametrizing the RG under the asynchronous hypothesis (*i.e.* only one component differs between two consecutive states). *Attractors -i.e.* terminal strongly connected components of the STG- are parts of the STG where the system stabilizes, interpreted as the long-term behavior of the system. There are two types of attractors : *steady states* and *cyclic attractors*, according to whether they are made up of one or several states.

Once the regulatory graph and its logical rules are defined, it is possible to run simulations specifying (or not) initial conditions using GINsim [25]. This software offers several functions such as the computation of all steady states, the reduction of the model based on a compressed STG, and various simulations according to pre-specified mutations. We have implemented a Python script that generates and extracts all the steady states of a model encoded in GINsim. Steady states are represented as rows of a Boolean matrix, whose columns correspond to the genes/proteins. The (i, j)-th coefficient of the matrix is equal to 1 if and only if the component j is expressed in the steady state i. An example is shown in Fig. 4.1.

a)



b)

Simulation 0	None (IL12 = 0)
Simulation 1	IL12 (IL12 = 1)

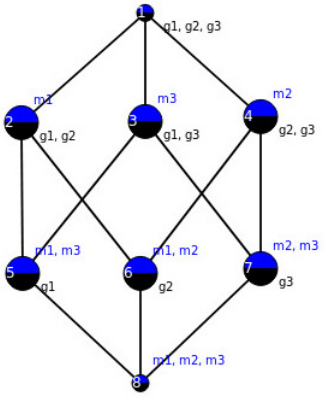
c)

Steady States	s_1	s_2	s_3	s_4
Simulation Condition	None	None IL12	None	IL12
IFNγ	0	0	1	1
IL4	0	1	0	0
IFNγR	0	0	0	0
IL4R	0	1	0	0
IL12R	0	0	0	1
STAT1	0	0	0	0
STAT6	0	1	0	0
STAT4	0	0	0	1
SOCS1	0	0	1	1
GATA3	0	1	0	0
Tbet	0	0	1	1

FIGURE 4.1 – **Small-scale network controlling the differentiation of Lymphocyte T helper (Th) with two input environments** (data extracted from [110]). (a) The network (has 11 internal compounds (gray nodes) and one external compound IL12 (green node)). Yellow labels give the logical function of each compound. Activations are represented by green arrows and inhibition by red arrows. (b) Input conditions used during simulation. The two values of the external node or input IL12 correspond to the stimulation or not of this gene.(c) Matrix crossing compounds and steady states with the conditions of stimulation (value 1 for an activated compound in a state, 0 otherwise) : the first condition of stimulation (IL12=0) generated the three steady-states s_1, s_2 and s_3 whereas s_2 and s_4 are the two steady-states which can be accessed with the condition IL12=1. In both cases, the convergence of the dynamics to one or the other steady-states depends on the initial state of the internal nodes.

	m_1	m_2	m_3
g_1	×		×
g_2	×	×	
g_3		×	×

(a) Formal context



(b) Concept lattice

FIGURE 4.2 – Example of a concept lattice. Table (a) describes a relation between the set of objects $G = \{g_1, g_2, g_3\}$ and the set of attributes $M = \{m_1, m_2, m_3\}$. Concepts are nodes in the graph (b) labelled with subsets of G and M, drawn with *LatViz* [111]. For instance, the top left is $(\{g_1, g_2\}, \{m_1\})$. Indeed, both g_1 and g_2 are in relation with m_1, and it is not possible to add g_3 ($g_3 \times m_1$ is lacking), m_2 ($g_1 \times m_2$ is lacking) or m_3 ($g_2 \times m_3$ is lacking). Similarly, the seven other formal concepts are $(\{g_1, g_2, g_3\}, \{\})$ -top-, $(\{g_2, g_3\}, \{m_2\})$, $(\{g_1, g_3\}, \{m_3\})$, $(\{g_2\}, \{m_1, m_2\})$, $(\{g_1\}, \{m_1, m_3\})$, $(\{g_3\}, \{m_2, m_3\})$ and $(\{\}, \{m_1, m_2, m_3\})$ -bottom-. They are organized in a lattice according to the set inclusion relationship between objects.

4.3.2 Formal Concept Analysis (FCA)

Our goal is to compare the steady states of a logical network over multiple simulations with different initial conditions. Moreover we are interested in an exhaustive enumeration of possible clusters of states without *a priori*. This involves, in particular, managing cluster overlaps.

Formal Concept Analysis (FCA) is a widely used data analysis technics that can be used for this purpose. In its most simple form, concepts formalize the duality extension and intension by extracting, from a binary relation between a set of objects and a set of attributes, the maximal subsets of objects that share the same subset of attributes [112]. Causality relations can be investigated within a lattice structure (Galois connection) by subconcept-superconcept relations. In bioinformatics, it has been used to derive phylogenetic relations among groups of organisms [113] and to exhibit clusters in large-scale interaction networks [114], [115]. In the following, objects are steady states of a Boolean network, and attributes are activations of biological compounds.

Formally, a data table allows building a context (G, M, I) where G (objects) and M (attributes) are two finite sets and $I \subset G \times M$ describes a relation between G and M. The set of attributes shared by all elements of $A \subset G$ is denoted by $A' = \{m \subset M | A \times \{m\} \subset I\}$. Similarly, the set of objects sharing all the elements of B is denoted by $B' = \{g \in G | \{g\} \times B \subset I\}$. The pair (A, B) is called a *formal concept* if $B = A'$ and $A = B'$, A being the extent and B the intent

of the concept. Additionally, the extent and the intent are closed sets, i :e A = A" and B = B". Equivalently and more intuitively, (A, B) is a formal concept precisely when every object in A is in relation with every attribute in B and it is not possible to add an element to A or B without breaking this property. For instance, for the relation shown in Fig.4.2, the concept lattice is a Boolean lattice with 2^3 nodes. The formal concept associated with m_1, is $(\{g_1, g_2\}, \{m_1\})$ and none of the elements g_3, m_2 and m_3 can be added to the concept. In this work, formal concepts identify all sets of states that share the same biological elements.

For a set of objects $A \subset G$, we associated to A an *attribute concept*, that is, the smallest formal concept which contains A, that is to be, (A'', A'). Dually, for a set of attributes $B \subset G$, the *object concept* is the largest formal concept which contains B, that is to be (B', B''). This allows us to identify either all the steady states sharing the same features, or all the biological elements characterizing a set of steady states. For the relation shown in Fig.4.2, there are six non-trivial formal concepts.

A partial order is defined over the family of formal concepts by $(A_1, B_1) \leq (A_2, B_2) \iff A_1 \subset A_2$ (or equivalently $B_2 \subset B_1$). The set of concepts forms a lattice. It means that every pair $((A_1, B_1), (A_2, B_2))$ of formal concepts has a greatest common subconcept $((A_1 \cap A_2), (B_1 \cup B_2)'')$, the *meet* and a lowest common superconcept $((A_1 \cup A_2)'', (B_1 \cap B_2))$, the *join*. The lattice can be represented as a graph. As shown in Fig.4.2b, formal concept are represented by a node in this graph. The top node is the concept with all objects and the bottom node is the concept with all attributes.

The lattice provides information about the impact of the addition or the deletion of objects or attributes over the classification process. More precisely, the structure of the concept lattice is suitable for capturing causalities through the use of implications and associations rules extracted from the lattice [109]. Without entering into details, we will implicitly use these causalities in the following section to explore the lattice associated with a Boolean network.

4.3.3 Building a lattice from a family of states in a Boolean Network

In the following, the term “compound" encompasses a gene, a protein or a biological compound. All are nodes in the regulatory network. In the following, the term “node" will refer to a formal concept in the lattice.

Let ϕ be a Boolean Network (BN) over a set of compounds $\boldsymbol{V}$ with values in $\{0, 1\}$. Let $\boldsymbol{S} \subset \{0, 1\}^{|\boldsymbol{V}|}$ be a family of steady states of ϕ (in general, it could be any subset of states). We define the relation I on $\boldsymbol{S} \times \boldsymbol{V}$ by setting that $(s, v) \in I$ if and only if v is activated in steady state s.

From context $(\boldsymbol{S}, \boldsymbol{V}, I)$ we build a lattice of formal concepts. As an example, Fig. 4.3 shows the formal context associated with 10 states of a BN. As shown in Fig. 4.3(b), a hierarchical clustering approach applied to the considered table highlights the role of the cluster

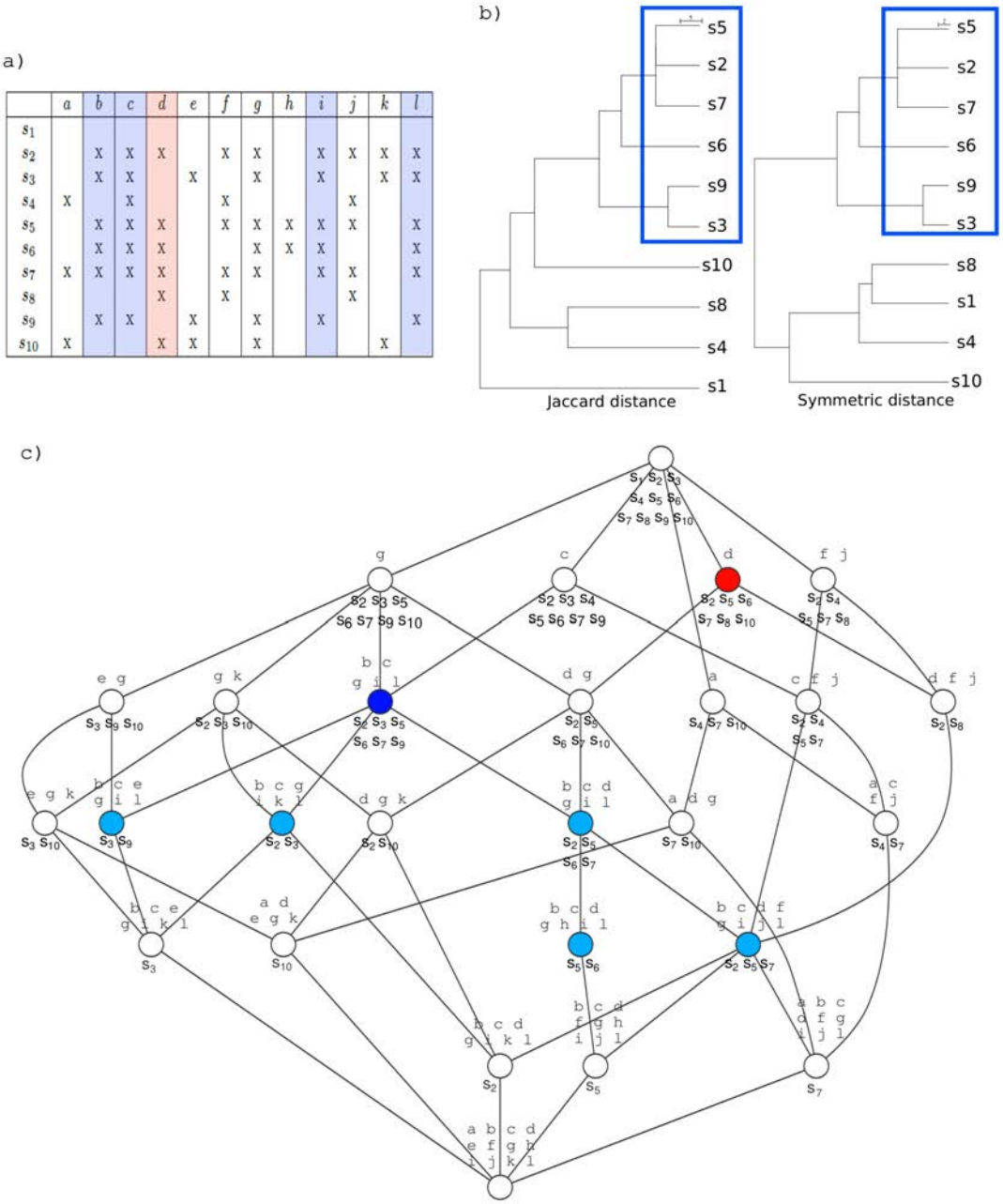


FIGURE 4.3 – **10-states matrix of a Boolean network and comparison between FCA and UGPMA clustering** (a) steady states are structured into a matrix where columns are compounds and rows states of the system. A compound is activated in a state if the corresponding cell in the matrix is checked (×). (b) Hierarchical clustering of states based on the Jaccard and symmetric distances. The blue frame characterizes the main common cluster, associated with the signature $\{b, c, g, i, l\}$. (c) Concept lattice derived from the matrix. The concept associated with $\{d\}$ is the red concept with objects $\{s_2, s_5, s_6, s_7, s_8, s_{10}\}$. These are the states for which the compound d is activated according to the table. The dark blue concept is associated with the signature $\{b, c, g, i, l\}$. It describes the family of six states for which the 5 compounds in the signature are activated. We notice that no formal concept contains the family of 4 compounds $\{b, c, i, l\}$. This means that the fifth compound g is always activated whenever the four compounds b, c, i, l are activated. and it should be added to the signature. There are five light blue concepts below the formal concept associated with $\{b, c, g, i, l\}$. They are called *variants* and can be automatically identified using the lattice structure.

$\{s_2, s_3, s_5, s_6, s_7, s_9\}$. A follow-up manual analysis suggests that the system's states are the states for which the compounds $\{b, c, g, i, l\}$ are simultaneously activated. On the contrary, the associated formal concept lattice in Fig.4.3(c) provides an exhaustive view of all existing relations between states and compounds. It contains 26 concepts corresponding to 26 subsets of objects which exhibit similar characteristics across their associated subset of attributes. This enables the reconstruction of associations. For instance, the red concept associated with $\{d\}$ contains states $\{s_2, s_5, s_6, s_7, s_8, s_{10}\}$: they are the states for which d is activated according to the table. Similarily, the lattice shows that the formal concept associated with $\{b\}$ (dark blue node in the lattice) has $\{b, i, l, c, g\}$ as set of attributes, which corresponds to the compounds identified above by post-processing the result of the clustering methods. Because there does not exist a formal concept (or node) with only $\{i\}$ or $\{l\}$, this means that b, i and l cannot be distinguished according to the relation provided by the table : they have similar columns. In the FCA formalism, we have there attribute equivalences (b is equivalent to i which is equivalent to l) and implication rules $b, g, i, l \implies c$ and $b, c, i, l \implies g$. Such an information, which is highly valuable for the study of dependencies in a network, cannot be obtained with clustering approaches and will be exploited all along our study.

As shown in this example, the main advantage of the lattice structure is to gather an exhaustive representation of the state families, and their inclusion relations, according to the activated compounds they have in common : instead of using a strong statistical signal it rather explores all dependencies between states and compounds, allowing these dependencies to be propagated along the lattice in order to investigate the role of the deletion or the addition of an activated compound in the clustering process.

4.4 Exploring the lattice of steady states according to biological signatures of phenotypes

4.4.1 Refinement of signatures according to phenotype knowledge

As stated in the introduction, it is common to interpret the different attractors of a model through one or several signatures, that is, sets of proteins or genes whose simultaneous activation is interpreted as a characteristic of a particular phenotype. In the following, we will denote by $\mathcal{S}g = \{v_1, ..., v_n\} \subset \boldsymbol{V}$ a phenotype signature, possibly provided by an expert.

A first added-value of FCA is to allow a systematic interpretation of the steady states with respect to a signature $\mathcal{S}g$. As the signature $\mathcal{S}g$ is a set of compounds, it can be associated to a unique nearest concept in the lattice, that is, the formal concept whose set of attributes contains all the biological compounds of $\mathcal{S}g$ and is minimal with respect to this property. As an example, let us assume that a phenotype is characterized by signature $\{b, c, i, l\}$. The lattice depicted in Fig.4.3(c) highlights that the greatest formal concept which contains these compounds as

attributes is $\{s_2, s_3, s_5, s_6, s_7, s_9\} \times \{b, c, g, i, l\}$. This concept conveys two informations. First, it points out to the six states that satisfy the signature. According to the notations introduced in Sec. 2.2, these states correspond to $\{b, c, i, l\}' = \{b, c, g, i, l\}'$. Second, it states that the activation of the compound g occurs whenever the 4 biological compounds in the signature are activated. These compounds correspond to $\{b, c, i, l\}$". As a matter of interpretation, this suggests that a refined signature, with the family of studied phenotypes (i.e., states), is $\{b, c, g, i, l\}$.

Based on this example, we define a refined signature of $\mathcal{S}g$ according to the set of steady states $\boldsymbol{s}$ to be the minimal set of attributes of the formal concept associated with $\mathcal{S}g$ in the lattice. With this definition, $\{b, c, g, i, l\}$ is the refined signature of any sub-family of $\{b, c, g, i, l\}$ which does not appear itself in the lattice or in another super-concept of the refined signature depicted in Fig.4.3(c). On the contrary, the refined signature of $\{c\}$ is $\{c\}$ itself, since the lattice contains the concept $\{s_2, s_3, s_4, s_5, s_6, s_7, s_9\} \times \{c\}$.

4.4.2 Variants

As explained above, the signature of a phenotype is the set of master genes or proteins characterizing this phenotype. Clearly, it may appear that several cells share the same master genes -and thus have the same canonical phenotype- although they differ by other “minor" (i.e. not master) components. Thus, the set of cells associated with a phenotype may contain subclasses, characterized by a subset of minor components. They constitute variants of the same canonical class. We extend the notion of refined signature to variants : we formally define the variants of a signature to be sets of attributes associated with concepts that are smaller than the concept associated with the refined signature and contain at least two states (to avoid signatures specific of a single state).

In the example depicted in Fig.4.3(c) for instance, starting from the biological signature $\{b, c, i, l\}$, we obtained the refined signature $\{b, c, g, i, l\}$, and this signature has five variants. The formal concept $\{b, c, e, g, i, l\} \times \{s_3, s_9\}$ allows stating that $\{b, c, e, g, i, l\}$ is the variant signature corresponding to the activated compounds shared by the states s_3 and s_9. Other variants are provided by concepts $\{b, c, g, i, k, l\} \times \{s_2, s_3\}$, $\{b, c, d, g, i, l\} \times \{s_2, s_5, s_6, s_7\}$, $\{b, c, d, g, h, i, l\} \times \{s_5, s_6\}$ and $\{b, c, d, f, g, i, j, l\} \times \{s_2, s_5, s_7\}$. This example illustrates that variants defined by the FCA framework correspond to all combinations of minor components shared by several phenotypes which satisfy the signature, and they eventually facilitate the understanding of the role of these additional markers. In Fig.4.3(b), we computed the supervised classification obtained with UGPMA clustering based on two metrics, Jaccard and symmetric distance for the 10 considered states. As shown in this clustering, the main signature $\{b, c, e, g, i, l\}$ can be easily identified (blue frame) and is related to one specific node of those hierarchical trees. Variants are the sub-clusters contained in this blue frame depicted by the nodes. We notice that three variants are identified by both approaches : $\{b, c, e, g, i, l\} \times \{s_3, s_9\}$, $\{b, c, d, g, i, l\} \times \{s_2, s_5, s_6, s_7\}$,

$\{b, c, d, f, g, i, j, l\} \times \{s_2, s_5, s_7\}$. They are all variants/clusters found in the UGPMA clustering. However, some variants like $\{b, c, g, i, k, l\} \times \{s_2, s_3\}$ and $\{b, c, d, g, h, i, l\} \times \{s_5, s_6\}$ can not be recovered in the hierarchical clustering. This illustrates the impact of the metric used in hierarchical clustering in terms of interpretation and classification. FCA uses a combinatorial approach rather than a statistical-based selection and performs a complete enumeration of possible variants for a given phenotype (or signature).

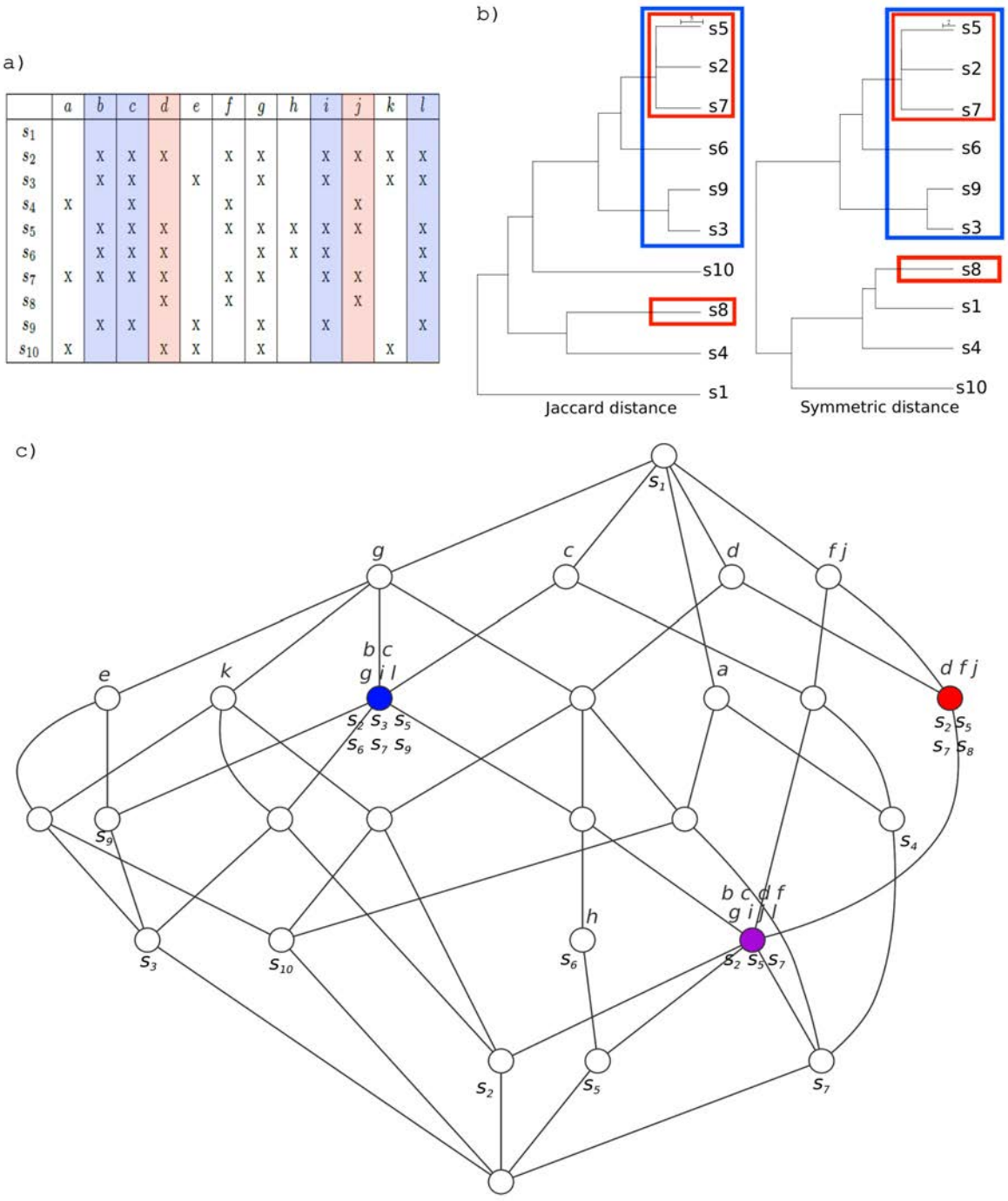
4.4.3 Identifying hybrids of several phenotypes characterized by their signatures

In this section, we will explain how FCA is suitable also for the analysis of sets of signatures. We assume that the steady states reached from different simulation conditions correspond to different canonical cell types, each associated with a biological signature characterized by master genes or proteins. According to the definitions introduced in the previous sections, each biological signature can be extended to a contextually-refined signature and can be enumerated.

One should notice that the definition of the refined-signature and the variants depend on the simulation conditions of the network. It has been observed in several studies that by modifying inputs (environments) and/or initial conditions, one may obtain steady states with more than one master gene [101], and possibly hybrid steady states for which several master genes associated with different cell-types signatures are activated. This corresponds to the concept of hybrid cell types in the literature [116], [117]. For instance, [116] showed that a sub-population of dendritic cells shared several surface markers with macrophages and appeared in the tumor microenvironment.

We define hybrid concepts to be concepts which are variants of at least two different cell-type signatures. We create a new hybrid cell type from two cell types t_1 and t_2 if and only if the meet $C_1 \wedge C_2$ of concepts C_1 and C_2, containing the refined signatures of t_1 and t_2, differs from the smallest concept in the lattice (bottom). The signature of the hybrid cell-type is the set of attributes of $C_1 \wedge C_2$. By construction, all genes and proteins in the signature belong to both t_1 and t_2 signatures. This concept itself may have variants, which are variants common to t_1 and t_2.

Fig.4.4 is an extension of the example shown in Fig.4.3. In this case, in addition to the cell-type defined by signature $\{b, c, i, l\}$, we consider another cell-type characterized by signature $\{d, j\}$. First, if we look at the UPGMA clustering [118], it appears that $\{d, j\}$ (red frame) is not identified by any of the metrics. Indeed, the steady state s_8 is isolated. However, the lattice shows that this second cell type has extended signature $\{d, f, j\}$ (red concept), with one variant, $\{b, c, d, f, g, i, j, l\}$ (purple concept). Importantly, this variant is also a variant of the first cell-type (blue concept). Therefore, both cell-types have an hybrid, with signature $\{b, c, d, f, g, i, j, l\}$. It differs from cell type $\{d, j\}$ by forcing the activation of $\{b, c, g, i, l\}$ and from cell type $\{b, c, i, l\}$



	a	b	c	d	e	f	g	h	i	j	k	l
s_1												
s_2		x	x	x		x	x		x	x	x	x
s_3		x	x		x		x		x		x	x
s_4	x		x			x				x		
s_5		x	x	x		x	x	x	x	x		x
s_6		x	x	x			x	x	x			x
s_7	x	x	x	x		x	x		x	x		x
s_8				x		x				x		
s_9		x	x		x		x		x			x
s_{10}	x			x	x		x				x	

FIGURE 4.4 – **Hybrid of two cell-types characterized by their signatures** (a) A Boolean matrix, the activated genes or proteins of the states of a Boolean network. A first phenotype is characterized by the biological signature $\{b, c, i, l\}$ (blue). Its refined signature is $\{b, c, g, i, l\}$, shared by states $\{s_2, s_3, s_5, s_6, s_7, s_9\}$. A second phenotype signature is characterized by $\{d, j\}$ (red). Its refined signature is $\{d, f, j\}$, shared by $\{s_2, s_5, s_7, s_8\}$. (b) Hierarchical clustering (average linkage) obtained from 2 distance matrices computed from the binary table (Jaccard and symmetric difference distance). The blue frame represents the set of states associated with the first signature and the red frame the set of steady states associated with the second signature. None of the metrics shows the link between s_8 and $\{s_2, s_5, s_7\}$. (c) Concept lattice associated with the matrix according to FCA. The concept associated with signature $\{b, c, i, l\}$ is in dark blue,the concept associated with the second signature $\{d, j\}$ is in red and the concept associated with the hybrid is in purple.

by forcing the activation of $\{d, f, j\}$. Interestingly, this cell-type has itself four variants which have no common compound with the hybrid and are called canonical variants to the cell-type $\{b, c, i, l\}$.

Together, the FCA framework allows the computation of all hybrid signatures for any family of signatures. Note that given a number of cell types n, the number of their variants can be exponential in n but the number of hybrids can be at most quadratic in n. In biological applications, hybrids are expected to be scarce and therefore may be easily analyzed manually.

4.4.4 Implementation

We implemented a python package *Foclass* to compute the concept lattice associated with a Boolean Network (available at `https://github.com/mwery/Foclass`). The Python package takes as input a Boolean Network together with simulation conditions formatted as a GINsim archive and a list of biological signatures provided as a text file. Those signatures are a set of activated genes associated with a biological phenotype as in [101]. The aim of the pipeline is to automatically classify all the steady states generated from the BN, according to the signatures and using the FCA.

The *ginsimToInputFile* command computes steady states for a given Boolean network. To that goal, the GINsim software (command line interface) computes all the steady states and the script generates the resulting matrix crossing states and compounds (genes or proteins). Each cell contains a Boolean value stating the presence of a compound in a state.

The analysisFCA command performs the FCA on the matrix : steady states are objects, compounds are attributes, null values are empty cells. The list of concepts is computed with a dedicated Python package and can be used to classify steady states. In order to improve the performances of the algorithm, the computation focuses on the list of concepts and omits the lists of relations within the lattice. This command analyses the set of formal concepts according to the classification of signatures. The family of input signatures is extended based on the network phenotypes provided by the GINSim simulation by selecting concepts with the smallest set of compounds containing this signature. Concepts associated with hybrid phenotypes are then listed, allowing for the classification of all steady states which either satisfy a single canonical signature or belong to hybrid phenotypes. The full exploration of the set of formal concepts also enables computing the number of variants – formal concepts that contain, at least, the signature in their attributes – for canonical and hybrid signatures. When the numbers of objects is less or equal than 300, the lattice can be drawn to show canonical and hybrid signatures.

4.5 Application to Th cells differentiation - Exhaustive and automatic study of hybrid phenotypes

We evaluate our tool on the Th cell differentiation process. Indeed, this well studied differentiation process generates a large number of different canonical cell types characterized by a (set of) master genes, proteins, or markers which constitute the cell type signatures. The different phenotypes have been proved to be associated with several conditions of simulations of logical networks of various complexities. They therefore provide a relevant case-study to demonstrate the added-value of FCA-based analyses.

4.5.1 Biological context

T helper cells are lymphocytes that mature in the thymus and play a central role in the adaptive immune system. There are several subsets of Th cells ; each type has been shown to express different cytokine profiles driving different immune response. Three main canonical cell types were first identified [119] : Th0 the naive form ; Th1 a pro-inflammatory type which expresses the specific transcription factor T-bet and produces IFN-γ ; and Th2, involved in allergic responses, which is induced by GATA3 expression and produces several interleukines (IL4, IL5). Over the last decade, several additional Th subtypes have been discovered : regulatory T cells (Treg), which depend on FOXP3 expression, and Th17 cells, induced by RORγT expression [120]. More recently, three additional subsets have been characterized [121], [122]. Th9 is linked to PU.1 expression and can be differentiated both from Th2 with stimulation of TGFβ and from Th0 with combination of TGFβ and IL-4. Th22 can be induced by stimulation of Th0 with TNF-α and IL-6 which drive STAT3 expression. Finally, T follicular helper cells (TFH) depend on several cytokine stimulations which cause BCL-6 expression.

4.5.2 Identification of variants in a small case study

Different models have been developed aiming at understanding the differentiation into each Th type under microenvironment change, and particularly in Boolean or multilevel framework [101], [102], [110]. As a first approach, [110] introduced a simplified model of the transcriptional regulatory network involved in the differentiation of $CD4^+$ naive T lymphocyte (Th) into Th1 or Th2, two active forms (Fig. 4.1(a)). Those two differentiated cell-types are triggered by two master genes (transcriptional factors) : Tbet for Th1 and GATA3 for Th2 (Fig. 4.5(a)). This network includes 12 Boolean internal components and one input component, IL12, which represents the cellular environment and is known to induce the differentiation of Th0 into Th1 or Th2 (Fig. 4.1(a)). Dynamical simulations can be performed for each possible value of input (IL12=0 and IL12=1) (Fig. 4.1(b)). They lead to that four different steady states, shown in Fig.

a)

Phenotypes	Associated signature
Th0	No protein = "empty"
Th1	Tbet
Th2	GATA3

b)

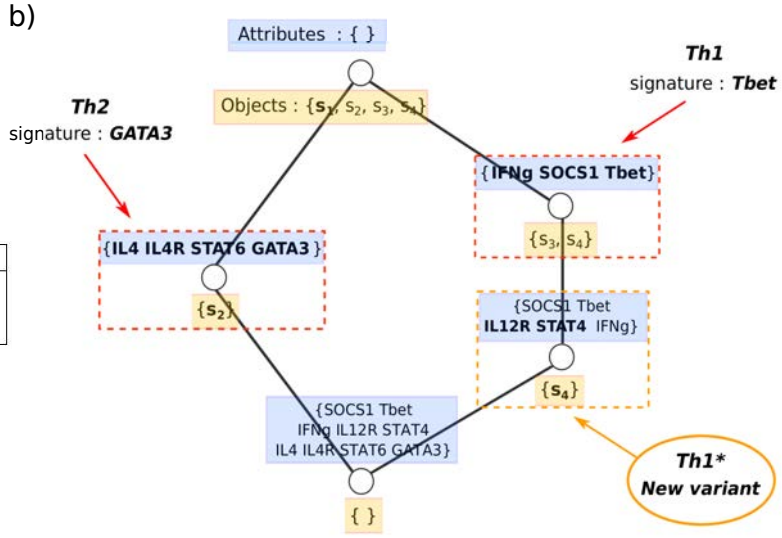


FIGURE 4.5 – **Concept lattice associated with the small-scale network controlling the differentiation of Lymphocyte T helper (Th).** This network was described in Fig. 4.1 together with the four steady states reachable in two simulation conditions. (a) Definition of the signature for each cell type. Th1 and Th2 are characterized by the expression of a single master regulator. The signature of Th0 is empty since it reflects the absence of expressed protein. (b) Concept lattice generated from the matrix with compounds (attributes) in blue frame and states (objects) in orange frame. The two red hatched rectangles represent the two formal concepts associated with the signatures of Th1 and Th2. The orange hatched rectangle shows the formal concept associated with the variant of Th1.

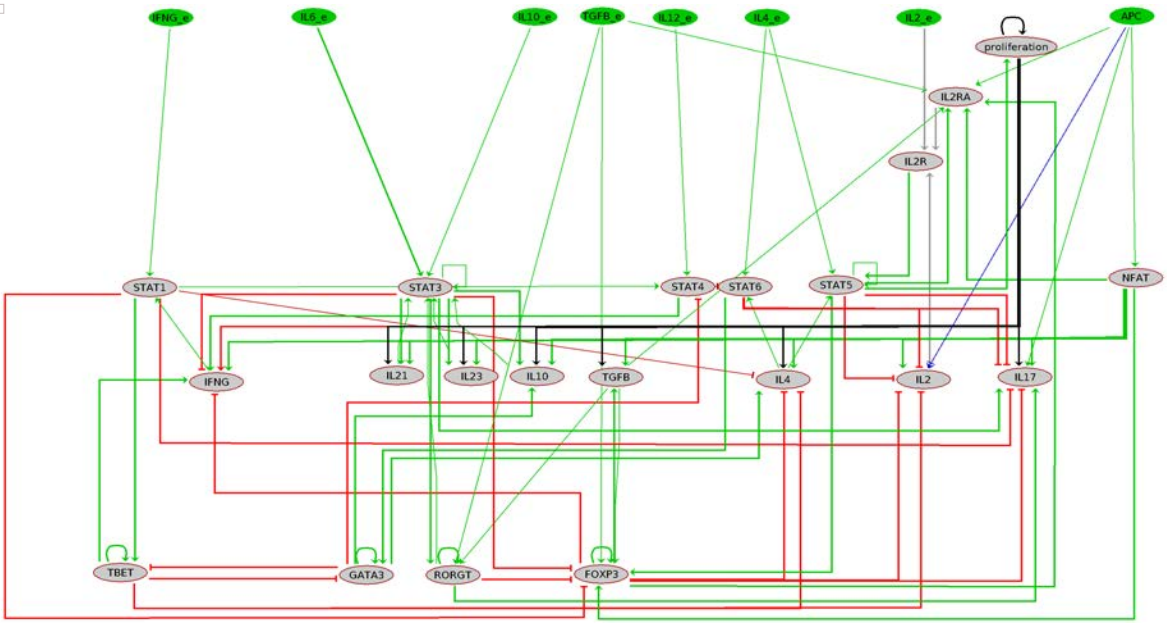
4.1(c) and in the associated concept lattice in Fig. 4.5(e). According to the lattice and the definitions introduced in the previous section the original signature of Th1, {Tbet}, can be extended by the refined signature {IFNγ, SocS1, Tbet}. In addition, this cell-type has a single variant, associated with refined signature {IFNγ, SocS1, Tbet, IL12R, STAT4}. This is in agreement with the biological role of the minor component STAT4 later confirmed in the literature. For instance, Thieu *et al* showed that STAT4 is required during the differentiation process of Th1 in order to achieve a complete phenotype [123]. On the left side of the lattice, the cell type Th2 (GATA3), associated with a single steady state has the refined signature {GATA3, IL4, IL4R, STAT6}.

4.5.3 Comparing the impact of different simulation conditions

Let us notice, however, that this network is no longer suitable to study the other subtypes (Treg and Th17). Indeed, the two transcriptional factors that regulate Treg and Th17 (FOXP3 and RORγT) are not involved. Also, the network does not take into account the influence of other external stimuli implied in Treg and Th17 regulation. A larger network was defined in order to understand the role of the microenvironment [101]. More precisely, Naldi *et al* [101] have integrated transcriptional pathways to enrich the model of Th differentiation. The extended model encompasses 65 components and is controlled by 13 inputs representing external environmental stimuli (see Fig. 4.6(a)). In their study, the authors tested several input combinations (see Fig. 4.6(b)) and evidenced that the dynamical simulations generated 38 steady states (see Fig. 4.6(d) - Condition 1). For each phenotype (cell type) a signature was introduced corresponding to the expression of one master regulator of the network (Tbet, GATA3, FOXP3 or RORγT for, respectively, Th1, Th2, Treg and Th17) (Fig. 4.6(c)). Moreover, the authors set up the initial state which represents the Th0 type. One of the main results of Naldi et al.'s publication was a classification table for the steady states of the system which was derived from the pre-determined signatures. With the table, subtypes of Th cells were introduced and characterized by sub-patterns of expressed proteins.

Our methods and tool allowed us to perform several simulations of Naldi's network [101]. First, we used the input conditions introduced in [101] (8 environment conditions described in Fig. 4.6(b)) and generated the 37 steady states that can be reached from the Th0 initial state. The lattice generated by the FCA contained 59 concepts (Fig. 4.7). Among them, we checked that all hybrids identified in [101] effectively corresponded to a hybrid concept as defined in our formalism. As expected and shown in Fig. 4.6(e), the 38 steady states could all be automatically characterized to satisfy either a canonical signature (4 steady states for Th0, 12 steady states for Th1, 6 steady states for Th2, 1 steady state for Th17 and 1 steady state for Treg) or a hybrid signature (4 steady states for Th1-Th17, 1 steady state for Th1-Treg, 3 steady states for Th2-Th17, 2 steady states for Th17-Treg, 2 steady states for Th1-Th17-Treg and 1 steady state

a)



b)

No Stimulation	None (All = 0)
APC only	APC (APC = 1)
Pro-Th1 (1)	APC, $IL2_e$, $IFNg_e$
Pro-Th1 (2)	APC, $IL12_e$
Pro-Th2	APC, $IL4_e$, $IL6_e$
Pro-Th17	APC, $IL6_e$, $TGFb_e$
Pro-Treg (1)	APC, $IL2_e$, $TGFb_e$
Pro-Treg (2)	APC, $IL10_e$

c)

Phenotypes	Signatures
Th0	"empty"
Th1	Tbet
Th2	GATA3
Th17	RORGT
Treg	FOXP3

d)

	Condition 1	Condition 2	Condition 3
Input	Fig 5b	Fig 5b	All
Initial State(s)	Th0	All	All
Nb Steady States	37	55	435
Nb Canonical Classes	5	5	5
Nb Hybrid Classes	6	7	7
Nb Formal Concepts	59	116	789
Lattice graph available	Fig 6a	Fig 6b	No

FIGURE 4.6 – **Network controlling the differentiation of Lymphocyte T helper (Th) with the input environments used for the dynamics simulation and the signatures for the classification.** Those data were described in [101]. (a) Reduced network of the differentiation of Th (35 components). Gray nodes correspond to internal compounds and green to the 13 inputs components. Activation regulations are represented with green arrows and indirect interactions resulting from the reduction are dotted arrows. Inhibitions are represented with red arrows. (b) Configuration of the input conditions used during simulation. Each row corresponds to one combination of inputs used for the simulation. (c) Initial signature for each cell type. Each row corresponds to one signature with the expression of one master regulators. The signature of Th0 is empty. However, some components in the system might be expressed. (d) Summary for the comparison of different simulation conditions of the network.

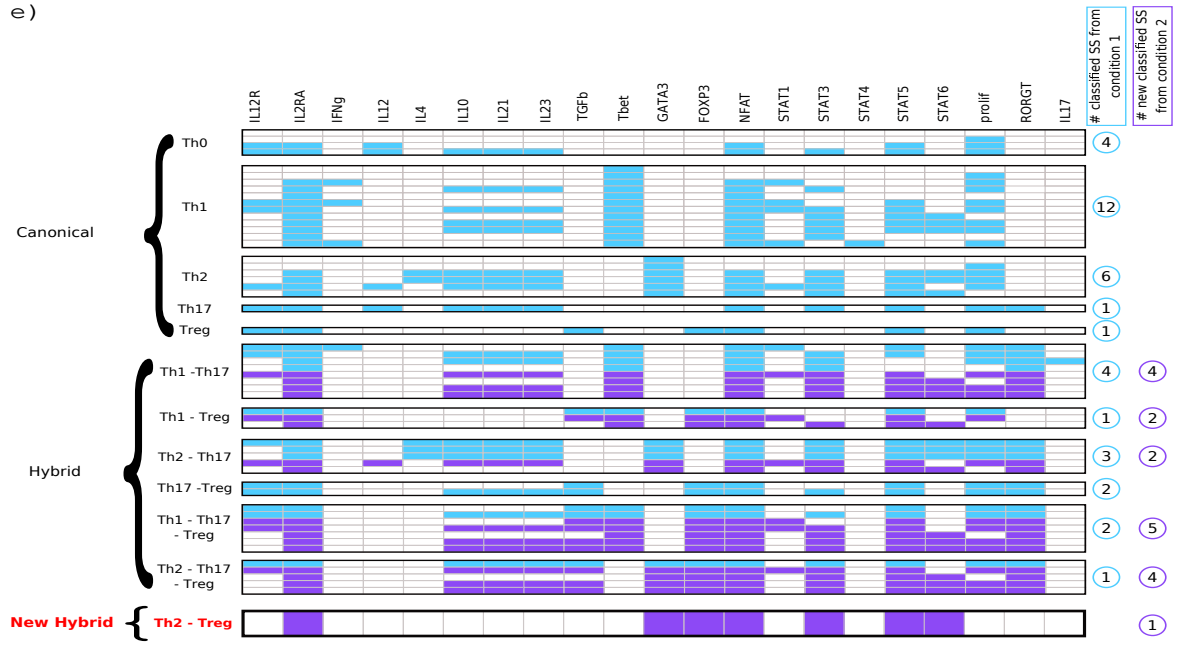


FIGURE 4.6 – **Network controlling the differentiation of Lymphocyte T helper (Th) with the input environments used for the dynamics simulation and the signatures for the classification.** (e) Comparison between the classified steady states from condition 1 (blue) and condition 2 (purple) of the network controlling the differentiation of Lymphocyte T helper (Th)

for Th2-Th17-Treg).

In order to study the influence of initial conditions on results, we simulated the network with the same input conditions (8 environmental stimuli) but by relaxing the constrain on the Th0 initial state and allowing any state to be considered as an initial state. This novel simulation condition generated 55 steady states (Fig.4.6(d) - Condition 2), whose analysis produced the concept lattice shown in Fig. 4.8. This lattice contained 116 formal concepts, illustrating the strong dependence of this type of study on the conditions of network simulation and the increasing complexity of studying it because the signature compounds are drowned in the entire lattice. Interestingly, in comparison with the previous simulation, almost all 17 additional steady states are associated with an already known hybrid class (Purple lines in Fig. 4.6(e)) : 4 novel steady states are associated with the hybrid cell type Th1-Th17, 2 with the hybrid cell-type Th1-Treg, 2 with the hybrid cell-type Th2-Th17, 5 with the hybrid Th1-Treg-Th17 and finallly 4 novel steady states with the hybrid cell-type Th2-Treg-Th17.

In contrast, one steady state could not be classified according to these canonical and hybrid cell-types. Based on our analysis, the hybrid type (Th2-Treg) is required to explain the data whereas there was no steady state associated with this phenotype in Naldi's work. Our prediction of this new hybrid has been in fact validated in the literature : Wang *et al* [124] worked on the role of $GATA3$ in the regulation of Treg function. The authors showed that the deletion of GATA3 expression induced an inflammatory disorder in mice with a decrease in FOXP3 expression. They also described that GATA3 can bind to a specific DNA sequence in the *FOXP3 locus* in the Treg cells. Our framework enables to understand better when this Th2-like Treg may occur. According to the simulations, this new hybrid phenotype appears only when the microenvironment is defined as pro-Th2, with activated APC, IL6 and IL4 input components.

A robust characterization of phenotypes

Comparing the two former simulation conditions highlights that extending the possibilities of initial states may have a large impact on the number of steady states and therefore on the lattice size, but it is only a low level vision of the cell behaviour. The number of hybrid for the other simulation condition does not change much : it increases only by 1 in our case. To push forward this idea and test the scalability of our method, we relaxed all the environmental conditions of [101] and simulated the network according to any initial state and any value for the environmental variables. This generated a family of 435 steady states (Fig. 4.6(d) - Condition 3). In this situation, the concept lattice contains 789 formal concepts. The graph is too large for being visualized with the *graphviz* package used in our method and we do not provide a figure for this lattice. However, our method evidenced that the numbers of canonical and hybrid classes remain the same as in the previous simulation, that is, 5 canonical classes and still 7 hybrid

classes. This demonstrates that our method is robust to the study of large environments and it allows to certify that all possible behaviors in terms of variants were described in the previous simulations.

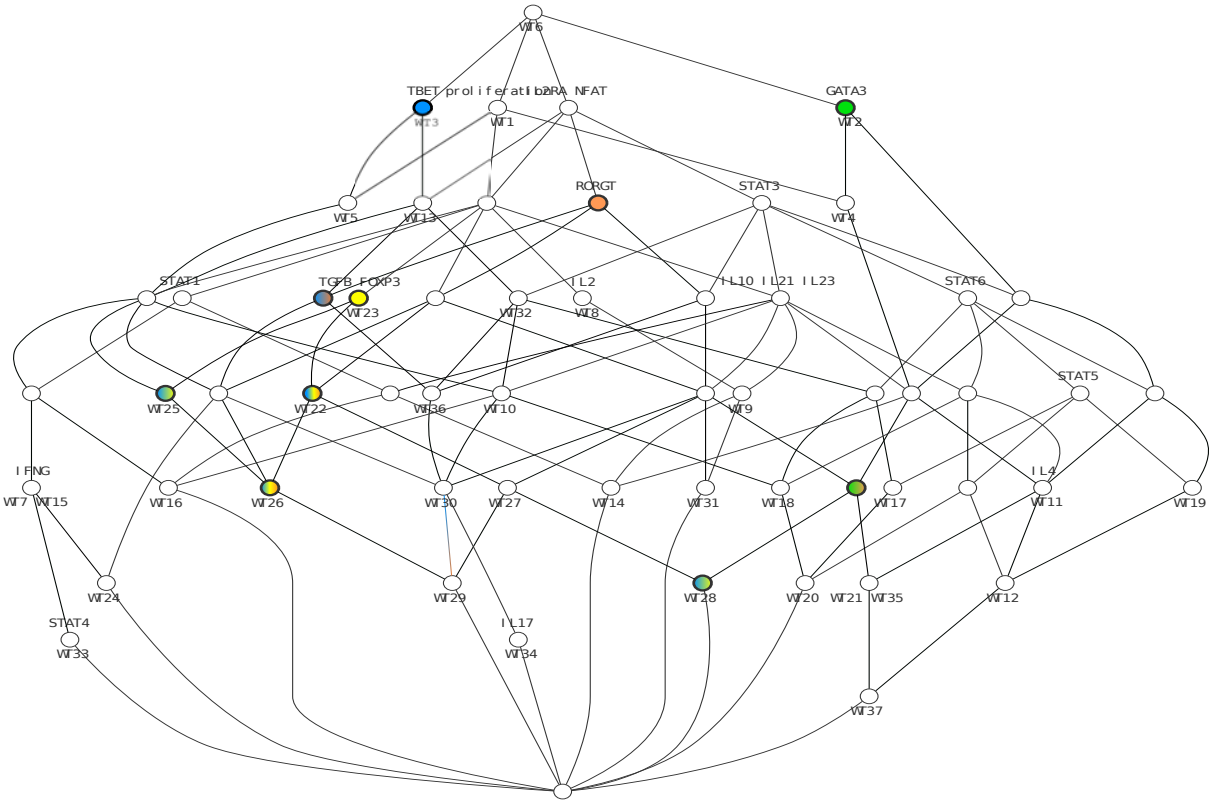


FIGURE 4.7 – **Lattice associated with a simulation of the network depicted in Fig. 4.6 with Th0 as initial state** The lattice is built with the 37 steady states obtained by simulating 8 different input environments with Th0 as initial state. Each node corresponds to a concept (59 nodes). Each edge represents the inclusion relation between two sets of states or components of the concepts. The five concepts associated with canonical signatures (plain) and the 6 concepts associated with hybrid signatures (gradient) are depicted in bold.

4.5.4 Classifying variants according to hybrids cell-types

Although the former analysis evidenced that the number of hybrid is relatively constant with respect to the simulation conditions of a Boolean network, the number of variants depends on the number of formal concepts and therefore may be intractable when either the simulation conditions or the number of components of the network increase. To address this issue, we advocate the use of hybrids as a way to classify variants. Fig. 4.9 details the classification of steady states for the simulation of Naldi's network [101] according to 8 stimuli/environmental conditions and Th0 as initial condition. In Fig. 4.9(a), the different variants are classified according to the few

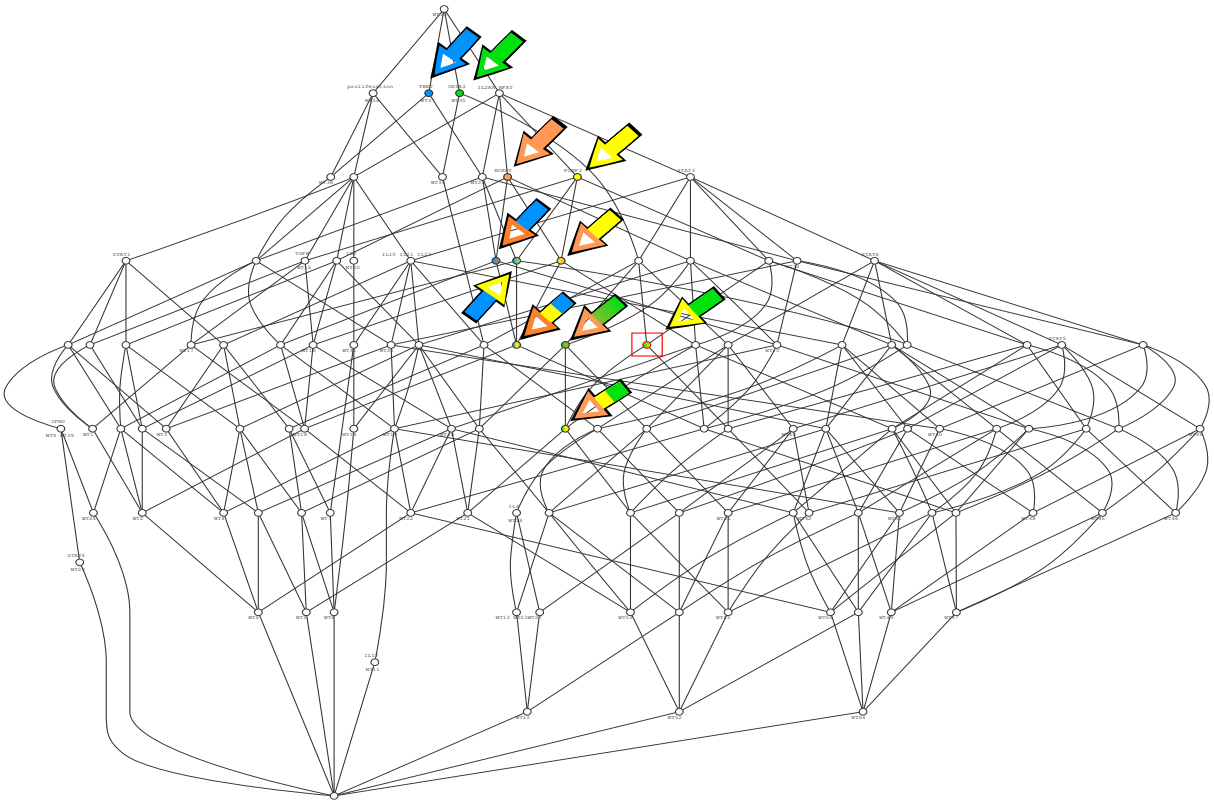


FIGURE 4.8 – **Lattice associated with a simulation of the network depicted in Fig. 4.6 with all possible values of internal genes or proteins as initial state.** The lattice is built according to the 55 steady states obtained by simulating 8 different input environments with all possible values of internal genes or proteins as initial state. It contains 116 nodes. In comparison with the lattice shown in Fig. 4.7, we notice that the number of concepts has nearly doubled, concepts associated with canonical (plain) and hybrid phenotypes (gradient) are rather stable, since a unique additional hybrid concept is created in the lattice (red frame).

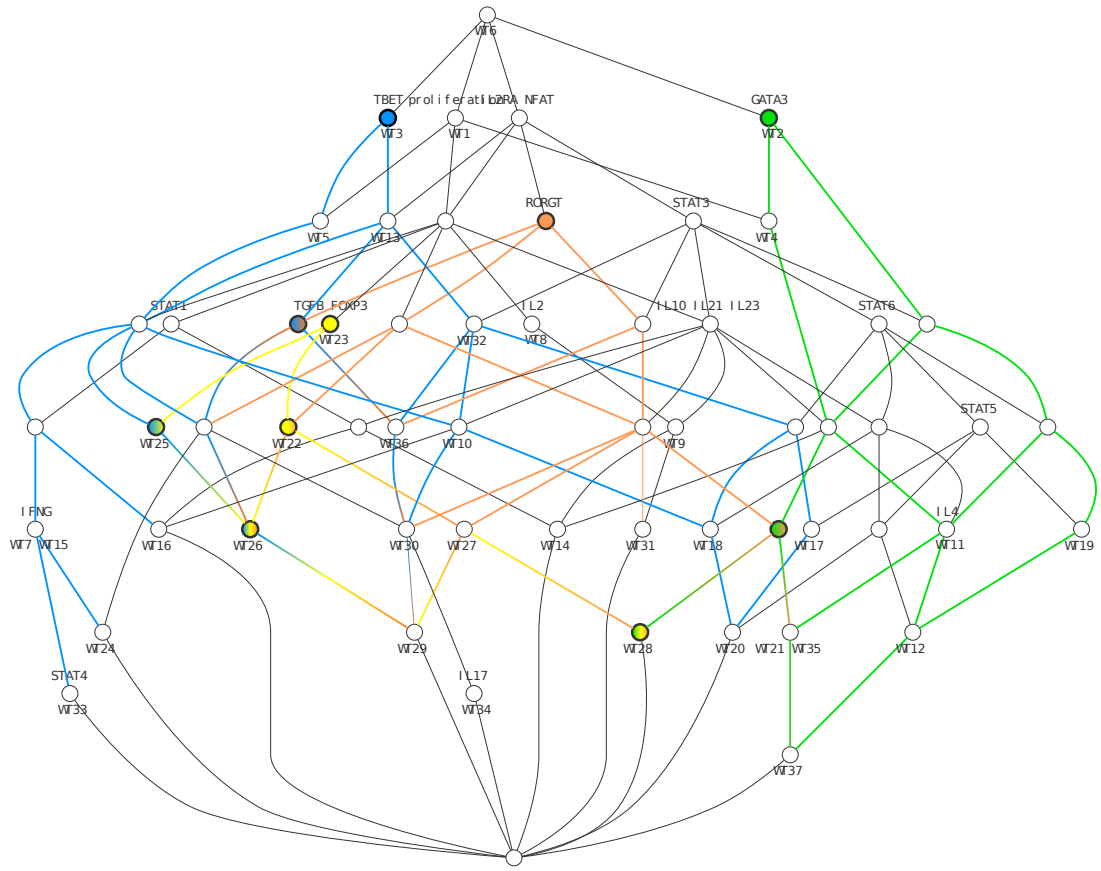
canonical or hybrid concepts they are linked to. For instance, the formal concept associated with Th1 signature (TBET) has 37 variants. Among them, two variants are hybrid nodes associated with Th1-Th17, Th1-Treg and Th1-Th17-TReg. 11 of the Th1 variants are actually Th1-T17 variants. 3 of the variants are Th1-Treg variants. Among them, 3 variants constitute the hybrid Th1-Th17-Treg and its variants. This analysis highlights that canonical concepts and hybrids allow the lattice to be decomposed into classes of variants such that each variant which is neither a hybrid nor a canonical concept belongs to a single class. Fig. 4.9(b) provides a synthetic representation of the variant classes and the global structure between them.

4.6 Discussion

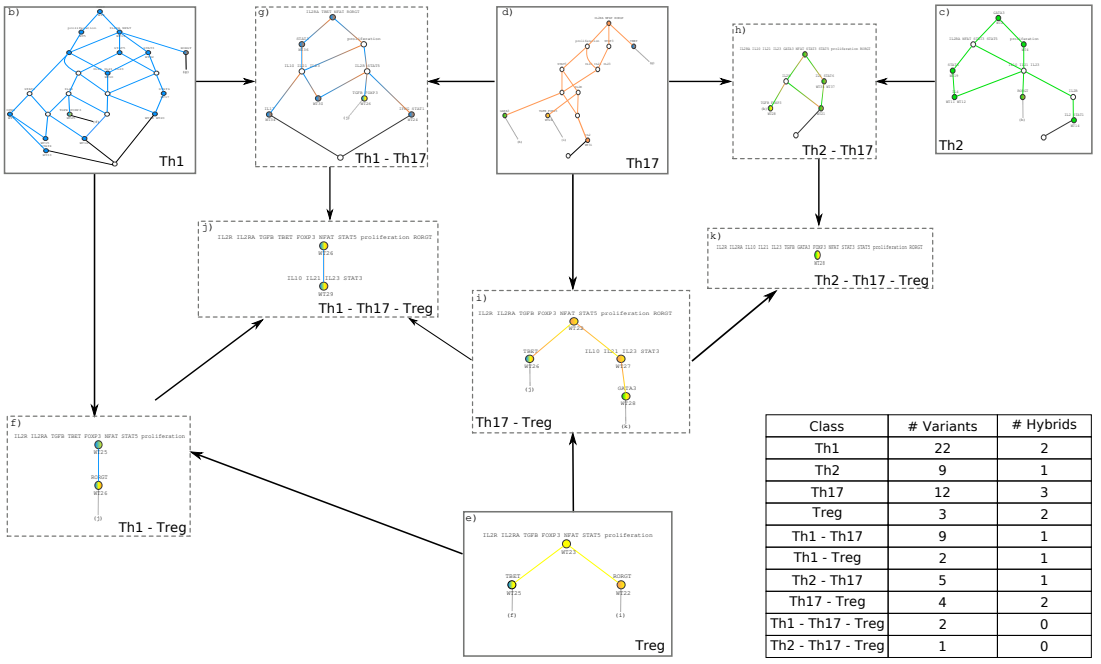
Data obtained from “omics" technologies provide a description of cellular compounds (gene, RNA, protein). Systems biology is based on this knowledge in order to analyze the dynamical behavior (phenotype) of the system in different conditions (specific environment or even mutations). When the number of simulation conditions increases, a major bottleneck of these analyses is the classification of the flood of steady states generated during the network simulation. Indeed, the main issue is that the system behavior is modeled by several phenotypes, each characterized by a few master regulators (genes, proteins or markers...). Distinguishing the activation of one or several master regulators in a large family of system's states is beyond the reach for standard clustering methods which are all subject to bias induced by their clustering metric.

To overcome this limitation, we promote the use of Formal Concept Analysis, a symbolic bi-clustering approach used in knowledge discovery and data mining. Our study suggests that FCA is accurate not only to extend expert-based signature according to the dependencies carried by the network dynamics, but also to automatically identify hybrid phenotypes associated with several signatures, together with initial conditions that may lead to new hybrid phenotypes. In addition, thanks to the hierarchy carried by the lattice structure of formal concepts, all the variants of canonical and hybrid phenotypes can be sorted in order to illustrate, for each phenotype, the role of biological compounds which are involved in signatures although they are not master genes. Such a distinction between master regulator and secondary regulator in signature was especially introduced in [98] ; our method may provide a structure to systematically investigate the role of secondary master regulators in variant phenotypes.

To illustrate our approach, we studied several Boolean models for the gene regulatory system controlling the differentiation of LTh [101], [110]. The identification of proteins involved in the transition from one phenotype to another are fully studied in cell plasticity or development of target therapy. In [101], the authors have specified one initial state for each simulation, the Th0 type because it is the naive, inactivated form of LTh. Several input sets have been used showing the microenvironment implication in the choice of sub-type differentiation. Our method resulted



(a)



Class	# Variants	# Hybrids
Th1	22	2
Th2	9	1
Th17	12	3
Treg	3	2
Th1 - Th17	9	1
Th1 - Treg	2	1
Th2 - Th17	5	1
Th17 - Treg	4	2
Th1 - Th17 - Treg	2	0
Th2 - Th17 - Treg	1	0

(b)

FIGURE 4.9 – **Analysis of the classification and identification of the hybrid classes based on the steady states generated from the condition 1 in Fig. 4.6(d)** (a) Lattice with each canonical class in plain color (Th1 ,Th2 ,Th17,Treg and hybrids classes. (b) Relation between each class based on the lattice. Each plain block is the sub-lattice of a canonical class. Each dashed block is the sub-lattice of a hybrid class. The nodes in blocks are the variants associated to the class.

in the same association of steady states according to the different subtypes of LTh as in [101], [110]. The classification in [101], [110] was done manually which needed to classify one steady state at a time. But in our approach, all steady states are taken in one time. Moreover, by relaxing the constraints of the stimuli conditions and the simulation initiation, we have rationalized and systematized the study of phenotypes and evidenced that a new hybrid should be added to the family of considered phenotypes to complete the set of possible subtypes of the system.

The first limitation of our method is related to performance when the Boolean Network's size increases. In terms of complexity, the number of formal concepts increases exponentially with respect to the number of objects (here, model steady-states) or attributes (model nodes). The computation of the lattice structure (subconcept-superconcept relation) is also computationally demanding, but we use this information only at the very last step of our workflow when computing the subgraphs associated with each signature. Thanks to this strategy of strictly limiting the computation of the lattice structure, our approach scaled to the study of all steady states on the full network from [101] (65 nodes with 24,267 steady states). Anyway, when too large, the number of steady-states or nodes of the model is a limitation for this method. Some specialized tools such as *In-close*[1] are helpful to handle a larger amounts of concepts.

Another level of complexity is related to the reachability properties. Dynamical simulations are performed given an initial state and some inputs, in order to identify reachable steady states and attractors. For instance, in [101], the initial configuration is required to study the plasticity of Th cell types after the classification of all steady states according to phenotypes' signatures. It provides a very useful information about which combinations of initial states and input components lead to a specific phenotype. With our method, the Th2-Treg hybrid class shown in Fig.6(e), was not identified in the seminal paper because it is generated only when the initial input is pro-Th2 and the initial state is different from Th0, which was not tested in [101]. However, our method is only able to store this information if the search space is constrained enough, as in [101]. Addressing this issue will require for instance to develop technics based on model-checking for the identification of initial configurations leading to any steady states. When the initial state information is not needed, the sole identification of steady-states is less computationally demanding.

A second limitation of this method is to define the signature considering only the presence (activation) of biological components. Clearly, signatures of phenotypes may also consist in the inactivation of some biological components, always missing in the concepts associated to canonical signatures. For instance, in [98], the Th1 signature is defined by the inactivation of all the master regulators but the secreted cytokines (IL4 and IL17). The reason of limiting signatures definition to activated components is the fact that FCA generates the concepts according to the presence of attributes shared by objects. A strategy to overcome this issue is to expand the matrix

1. `https://sourceforge.net/projects/inclose/`

used as input for FCA by duplicating each attribute (compound) v in v, *not* v with the implicit constraint that exactly one of them is present at a time : v + *not* $v = 1$. A complementary enrichment of the method could be to take into account continuous – or at least multiple – values, either derived from differential equations modelling, or corresponding to biological data with samples as object and the gene/protein as attribute. This extension could be done be relying on FCA tools that handle numerical tables.

Our method has been developed for the analysis of steady states. As described in Sec. 4.3.1, Boolean Networks also generate *cyclic attractors*, describing oscillations through several states of the system. An interesting extension is to consider these cyclic attractors, and associate to them one or several signatures. The identification of cyclic attractors is not an easy task. Moreover, they are composed of a -possibly very large- set of states, and it is not always obvious to express them in a compact way. To cope with these difficulties, some strategies are possible. Usually, we try to express the cyclic attractors by one or several "schemes". They are defined with abstract states with constant components (meaning that all the states gathered in the scheme have this component fixed to the same value, 0 or 1), and cycling components. If the stabilized components match a signature, we may associate this signature to the attractor.

Finally this method provides a way to measure to what extend a perturbation of the model (i.e. a mutation simulated by blocking a node to value 0 or 1) affects the system. Indeed, we can compare the signatures of the stable states obtained in the wild-type and a mutant model. Considering the mutation of a node of the network, its impact may be measured through a sort of robustness, depending on if the attractors obtained in the mutant model are assigned to the same phenotypes/signatures as the attractors of the wild-type (although different). Hence, a mutation may affect the dynamics in terms of loss or gain of phenotypes, loss or gain of reachability, etc...(e.g., loss of tumors suppressor in cancer cells. The impact of mutations on biological systems represented with logical networks was highlighted for instance in [98], [100]). Intuitively, we expect that the mutation of a master gene regulator in the signature of a phenotype will strongly impact the model. The definition of signatures in the context of gene deletion or ectopic still deserves to be further studied. A perspective is to formally model the effect of a perturbation as an operation over the concept lattice. The approach would require to compute the steady-states for the wild-type model and all the perturbed models to consider (for instance, models perturbed with at most two knock-out or ectopic perturbations). The modelling issue would then be to figure out how signature and hybrid can be defined on formal concepts of this extended set of steady-states, by taking into account the type of perturbations that was performed to generate the considered steady states.

En conclusion, la méthode présentée permet une classification automatique des états stables générés à partir d'un réseau booléen. Cette classification est réalisée par analyse par concept formel et se base sur l'activation de certains noeuds, définis comme une signature pour chaque phénotype. De plus, elle permet l'identification de nouveaux phénotypes hybrides des lymphocytes T. En complément de cette classification, les signatures des phénotypes sont enrichies grâce aux dépendances de régulation du système. Les éléments supplémentaires peuvent jouer un rôle causal dans l'expression des entités constituant la signature et tout comme le comportement que va prendre le système biologique.

Conclusion & Perspectives

I Conclusion

En conclusion de ce travail de recherche, plusieurs contributions ont été apportées dans le but de proposer une nouvelle classification des entités (individus et états stables). De plus, la recherche de causalité dans les données -omiques et les connaissances favoriserait, à plus long terme, le développement de nouvelles cibles thérapeutiques.

La première contribution concerne des connaissances générales de la biologie médicale. Plusieurs types de signatures ont été recensés et ces catégories sont spécifiques à des comparaisons de populations et d'analyses statistiques.

Les contributions méthodologiques que nous avons produites sont les suivantes.

Dans le chapitre 2, nous avons introduit une méthode utilisant les technologies du Web sémantique (RDF) pour structurer et stocker des données -omiques. Elle vise à être la plus générique possible et permettre l'intégration d'autres couches -omiques dans le futur. La structure n'est pas uniquement orientée sur le patient mais aussi sur le gène. Le chapitre 3 présente une requête SPARQL complexe qui a été construite à partir d'une définition formelle. Le temps de calcul a été amélioré en passant d'une requête linéaire à une division en 2 sous-requêtes. Le filtrage se basant sur un premier ensemble de contraintes a permis l'optimisation de l'interrogation de ce triplestore.

Par ailleurs, dans le chapitre 4, nous présentons une méthode qui classe automatiquement des états stables en combinant l'analyse par concept formel (FCA) avec la dynamique d'un réseau booléen biologique. L'intérêt de la classification par FCA, comparée aux analyses classiques de clustering non supervisé, est qu'il s'agit d'une approche exhaustive. En effet, les concepts représentent l'ensemble des classifications possibles alors que les approches classiques comme le clustering hiérarchique ne proposent qu'un seul classement où chaque élément ne peut appartenir qu'à un seul cluster. C'est grâce à cet aspect que les classes hybrides ont pu être obtenues en FCA et pas en clustering hiérarchique.

Enfin, la principale contribution au domaine de la bioinformatique que nous avons produite est une approche automatique pour individualiser les données -omiques d'une étude clinique. Dans la continuité du chapitre 2, nous avons défini au chapitre 3 de manière formelle le concept d'*expression Individually-Consistent Trait Loci* (eICTLs) en combinant des données -omiques hétérogènes. Ces éléments identifient un lien de causalité potentiel entre la génomique et la

transcriptomique et réduisent la dimensionnalité des données -omiques classiques. Ces entités sont identifiées comme étant spécifiques à certains patients. Plusieurs entités pourraient être partagées par un sous-ensemble de patients.

Une autre contribution bioinformatique, présentée au chapitre 4, est une définition formelle d'une signature phénotypique dans un contexte de réseau booléen. De plus, la classification automatique des états stables a mis en évidence des variants, qui sont des dérivés des signatures canoniques, et des phénotypes hybrides qui sont des cas particuliers de variants.

De la même manière que pour le chapitre 4, la FCA pourrait être utilisée pour identifier des groupes d'eICTLs partagés par plusieurs patients et qui pourrait être une signature causale potentielle pour cet ensemble de patients.

Le SLE est connu comme étant une maladie hétérogène dans ses symptômes mais aussi complexe dans ses mécanismes d'apparition. Toutes les analyses classiques ont identifié des signatures composées principalement de gènes liés aux interférons (IFN). Comme l'a montré le chapitre 1, l'utilisation des critères cliniques comme critères de stratification ne permet pas de calculer des signatures fiables. La recherche de eICTLs permet l'identification de signaux faibles associés au SLE et spécifiques à certains patients. A terme, ces éléments peuvent être regroupés en sous-ensembles qui caractérisent des groupes de patients. En utilisant la même approche de classification qu'au chapitre 4, la FCA énumérerait tous les concepts pour lesquels un ensemble de eICTLs caractérisent un ensemble de patients. Ces différents regroupements pourraient être des signatures causales pathologiques du SLE et servir comme nouvelle stratification des patients SLE. Des éléments de ces signatures pourraient être utilisés à long terme comme de nouvelles cibles thérapeutiques.

II Perspectives

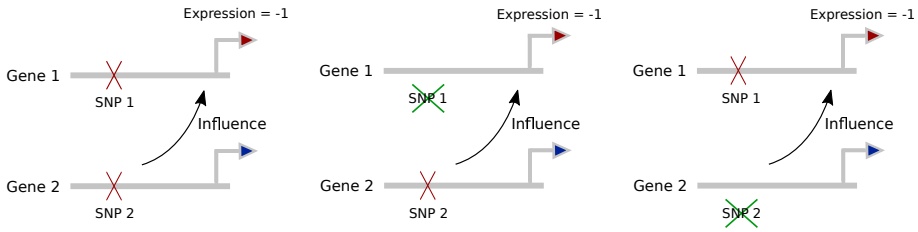
II.i Identifier des influences causales indirectes

La définition présentée dans le chapitre 3 permet d'extraire une influence directe entre la présence d'un SNP est la variation d'expression de son gène.

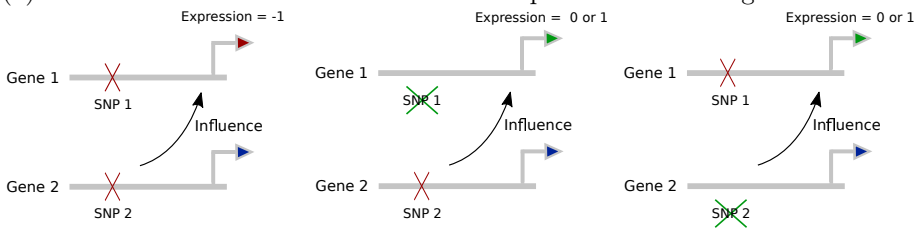
Cependant cette influence peut être éloignée, indirecte et ce pour diverses raisons. Soit le SNP n'est pas dans la séquence définie pour le gène, soit il peut s'agir d'un SNP ayant une influence sur un autre gène. Dans ce cas, ce lien ne se ferait pas sur l'expression de ce deuxième gène mais sur sa fonction par exemple ou son repliement.

Les maladies auto-immunes sont définies comme des maladies multi-factorielles. Autrement dit, leur déclenchement serait du au cumul de plusieurs facteurs. Ce phénomène cumulatif peut se retrouver aussi dans la recherche des liens indirects où l'interaction de deux SNPs peut influencer l'expression d'un gène. Deux types de combinaisons peuvent être envisagés et sont décrits

dans la figure 4.10.



(a) Influence dominante entre les deux SNPs présents sur deux gènes différents



(b) Influence récessive entre les deux SNPs présents sur deux gènes différents

FIGURE 4.10 – **Représentation schématique des conditions pour l'identification des influences indirectes candidates.** (a) La condition est qu'au moins un SNP ($SNP1\ OR\ SNP2$) soit présent sur l'un des deux gènes pour entraîner la diminution de l'expression du gène 1 par rapport à la population contrôle. (b) La condition est que les deux SNPs ($SNP1\ AND\ SNP2$) soient présents sur les deux gènes pour entraîner la diminution de l'expression du gène 1 par rapport à la population contrôle.

La première combinaison est représentée par une porte logique OU. Concrètement, il existe deux SNPs : SNP 1 est localisé sur le gène 1 et SNP 2 est localisé sur le gène 2. Lorsqu'au moins un des deux SNPs est présent, cela a une influence sur l'expression du gène 1. On parlera alors d'influence dominante.

La seconde combinaison est caractérisée par une porte logique ET. En gardant la description des deux SNPs, il est nécessaire que les 2 SNPs soient présents pour observer une variation de l'expression du gène 1. Dans ce cas, il s'agit d'une influence récessive.

La recherche de ces influences indirectes serait possible par l'interrogation du triplestore via la construction de requêtes SPARQL. Cependant, les premiers essais de ce type de requêtes n'ont pas réussi le passage à l'échelle. En effet, rechercher ces liens indirects va drastiquement augmenter l'espace des recherches. Une première optimisation a été tentée afin de n'interroger que les SNPs qui étaient sur les mêmes chromosomes mais le temps de calcul restait beaucoup trop long. D'autres pistes seraient d'ajouter des contraintes sur les requêtes ou d'optimiser la structure de celles-ci.

II.ii Enrichissement du modèle transomique par intégration d'autres données

Intégration d'autres données -omiques de patients - Dans le cadre de la thèse, des données de génotypage et de transcriptomique étaient mises à disposition. Dans l'IMI PRECISESADS, des données de métabolomique ont aussi été mesurées. Le modèle proposé dans le chapitre 2 pourrait être étendu afin d'intégrer ce type de données mais aussi des données de protéomique par exemple. A plus long terme, il pourrait être suffisamment complexe pour intégrer tous les types de données. Le schéma RDF évoluerait en même temps que de nouvelles couches seraient intégrées, mais l'alimentation centrée sur le patient n'aurait que peu ou pas de changement. Ce modèle plus complet permettrait l'exploration de l'ensemble des données que la technologie actuelle et future peut nous offrir. Cela aura pour conséquence d'augmenter les connexions entre les couches -omiques mais aussi de diversifier les questions biologiques associées à ce modèle. Son interrogation va complexifier les requêtes SPARQL mais pourrait optimiser le temps d'exécution. En effet, de nouvelles contraintes vont s'ajouter pour chaque couche.

Intégration de connaissance - Les connaissances contenues dans les bases de données externes comme KEGG, REACTOME ou Uniprot apporteraient des informations importantes sur la fonctionnalité des gènes. Il n'est bien sûr pas envisageable d'intégrer localement toutes ses bases de données. De nombreuses bases sont désormais accessibles sous forme de triples-tore. L'utilisation de requêtes fédérées va permettre l'interrogation de ces bases de données. Les connaissances extraites seront ajoutés sous forme de nouvelles contraintes pour les requêtes locales.

Intégration de données issues du single-cell - L'évolution de la technologie en biologie a conduit à la mesure des différentes couches -omiques à l'échelle de la cellule (single-cell). Ainsi, lors de ce type d'analyse, les mesures sont faites pour chaque individu et pour chaque cellule.

L'intégration de ce type de données dans le schéma ne va pas beaucoup changer sa structure générale. En effet, une seule entité sera ajoutée, celle du type cellulaire. En revanche, l'alimentation et l'interrogation de ce nouveau modèle subirait de plus grands changements. Tout d'abord, le fait de discrétiser les données d'expression poserait la question de la référence. L'intervalle de référence serait-il calculé par rapport à tous les types cellulaires des individus sains ou par type cellulaire chez les individus sains ? Ensuite, la recherche des eICTLs serait-elle possible si les données ne sont plus spécifiques qu'aux patients ? Enfin, pourrait-on combiner des données issues d'analyses single-cell avec des études plus classique ?

II.iii Combiner les signatures transomiques et les analyses des connaissances

Les chapitres 2 et 4 présentent deux méthodes différentes pour permettre l'identification de causalité entre plusieurs entités. La première est basée sur une analyse par raisonnement sur les

données et les liens qui existent entre elles. La seconde réalise une analyse sur les connaissances des interactions entre les éléments biologiques.

Lors de cette thèse, une autre approche a été envisagée afin de combiner l'utilisation des données avec les réseaux de régulation de gènes tout en gardant l'aspect centré sur le patient. L'hypothèse est qu'une signature dépend de caractéristiques spécifiques aux individus ainsi que de dépendances liées aux régulations biologiques.

L'objectif général était d'utiliser un même réseau biologique pour tous puis de l'individualiser par patient. L'ensemble des eICTLs spécifiques d'un individu serait appliqué au système en définissant l'état initial du réseau. Par exemple, pour un patient et un gène donné, si le patient possède un SNP sur le gène et que l'expression du gène a pour valeur -1 alors la valeur du gène dans le réseau du patient est fixée à -1. En revanche si le gène ne possède pas de SNP ou si la valeur d'expression est de 0, la valeur du noeud n'est pas fixée. En appliquant le principe des ensembles minimaux d'intervention (MIS), certains noeuds du réseau vont trouver leur valeur fixée en plus des eICTLs . La simulation dynamique du système propage le signal jusqu'à atteindre un état stable reflétant les expressions des gènes de ce même individu. Pour réaliser cette approche, le logiciel *Caspo*, et plus particulièrement sa fonction *control*, identifierait ces noeuds supplémentaires fixés. Ces derniers permettraient ainsi d'expliquer les expressions observées en fonction des eICTLs et ce pour chaque individu. Ces éléments seraient aussi utilisés pour définir un ensemble de causes possibles pour chaque individu. Ceci amène à la définition d'une signature comme étant l'ensemble des interventions communes à plusieurs individus.

Plusieurs limites relatives au réseau utilisé sont apparues, ne me permettant pas de poursuivre cette approche pour le moment. Au vu des données disponibles, l'outil nécessite un réseau qui soit booléen et qui contienne suffisamment de noeuds pour être exploitable. Cette méthode se voulant la plus généraliste possible, l'extraction du réseau devait se faire sans *a priori* à partir d'une liste de gènes différentiellement exprimés dans des bases de données déjà existantes comme PathwayCommons. Plusieurs biais existent dans l'utilisation des bases de données. Tout d'abord, beaucoup de gènes ne sont pas retrouvés car les identifiants utilisés ne sont pas forcément les mêmes dans toutes les bases de données. Ensuite, certains gènes et voies sont plus étudiés que d'autres et cela se ressent lors de l'extraction, puisque ces éléments sont beaucoup plus annotés, reliés. Enfin, même si l'extraction est une réussite, il faut apprendre les règles logiques pour chacun de ces noeuds. Or, cet apprentissage est compliqué car peu de bases de données annotent ce type d'information et l'extraction de ces règles par une analyse bibliographique est une étape très contraignante.

II.iv Recherche des eICTLs dans une autre maladie complexe : le cancer

Le modèle transomique a été appliqué à une maladie auto-immune complexe, le SLE. D'autres types de pathologie peuvent être pris en compte à partir du moment où les données disponibles

sont du même type.

L'interrogation du modèle pour la recherche des eICTLs peut aussi être transposable à ces pathologies. Cependant, il existe un cas particulier pour lequel la définition des eICTLs s'applique difficilement. Il s'agit du cancer. L'identification de couples SNP-gène dans un contexte pathologique permettrait d'obtenir une signature contenant des éléments pouvant être liés à la cause de la pathologie.

L'hypothèse avancée est que les mutations sont présentes avant le déclenchement de la pathologie et entraîne à long terme le dysfonctionnement cellulaire. Les cellules cancéreuses sont caractérisées par une prolifération anarchique. Cela entraîne une instabilité génétique ayant pour conséquence la génération de mutation au cours de la progression tumorale. Dans ce cas, les mutations ne sont plus considérées uniquement comme de potentielles causes de la maladie mais aussi comme des conséquences.

Deux options sont donc envisagées pour pallier ce problème. Soit la méthode ne peut s'appliquer aux études cancéreuses, soit il faut considérer que les eICTLs candidats sont un mélange de causes et de facteurs d'aggravation de la pathologie. Une expertise serait donc nécessaire afin de différencier ces deux états.

BIBLIOGRAPHIE

[1] S. T. Ngo, F. J. Steyn et P. A. McCombe, « Gender differences in autoimmune disease », *Front Neuroendocrinol*, t. 35, n° 3, p. 347-369, août 2014.

[2] P. Invernizzi, S. Pasini, C. Selmi, M. E. Gershwin et M. Podda, « Female predominance and X chromosome defects in autoimmune diseases », *J Autoimmun*, t. 33, n° 1, p. 12-16, août 2009.

[3] V. K. Kuchroo, P. S. Ohashi, R. B. Sartor et C. G. Vinuesa, « Dysregulation of immune homeostasis in autoimmune diseases », *Nat Med*, t. 18, n° 1, p. 42-47, jan. 2012.

[4] S. V. Ramagopalan, D. A. Dyment, W. Valdar, B. M. Herrera, M. Criscuoli, I. M. Yee, A. D. Sadovnick et G. C. Ebers, « Autoimmune disease in families with multiple sclerosis : a population-based study », *Lancet Neurol*, t. 6, n° 7, p. 604-610, juil. 2007.

[5] A. Parkkola, A. P. Laine, M. Karhunen, T. H ?rk ?nen, S. J. Ryh ?nen, J. Ilonen et M. Knip, « HLA and non-HLA genes and familial predisposition to autoimmune diseases in families with a child affected by type 1 diabetes », *PLoS One*, t. 12, n° 11, e0188402, 2017.

[6] M. P ?rez-De-Lis, S. Retamozo, A. Flores-Ch ?vez, B. Kostov, R. Perez-Alvarez, P. Brito-Zer ?n et M. Ramos-Casals, « Autoimmune diseases induced by biological agents. A review of 12,731 cases (BIOGEAS Registry) », *Expert Opin Drug Saf*, t. 16, n° 11, p. 1255-1271, nov. 2017.

[7] H. Xu, M. Liu, J. Cao, X. Li, D. Fan, Y. Xia, X. Lu, J. Li, D. Ju et H. Zhao, « The Dynamic Interplay between the Gut Microbiota and Autoimmune Diseases », *J Immunol Res*, t. 2019, p. 7 546 047, 2019.

[8] K. S. Lang, A. Burow, M. Kurrer, P. A. Lang et M. Recher, « The role of the innate immune response in autoimmune disease », *J Autoimmun*, t. 29, n° 4, p. 206-212, déc. 2007.

[9] J. Suurmond et B. Diamond, « Autoantibodies in systemic autoimmune diseases : specificity and pathogenicity », *J Clin Invest*, t. 125, n° 6, p. 2194-2202, juin 2015.

[10] C. Castro et M. Gourley, « Diagnostic testing and interpretation of tests for autoimmunity », *J Allergy Clin Immunol*, t. 125, n° 2 Suppl 2, S238-247, fév. 2010.

[11] J. H. Cho et M. Feldman, « Heterogeneity of autoimmune diseases : pathophysiologic insights from genetics and implications for new therapies », *Nat Med*, t. 21, n° 7, p. 730-738, juil. 2015.

[12] A. Li et R. C. Bergan, « Clinical trial design : Past, present, and future in the context of big data and precision medicine », *Cancer*, sept. 2020.

[13] B. Lee, S. Zhang, A. Poleksic et L. Xie, « Heterogeneous Multi-Layered Network Model for Omics Data Integration and Analysis », *Front Genet*, t. 10, p. 1381, 2019.

[14] Z. D. Stephens, S. Y. Lee, F. Faghri, R. H. Campbell, C. Zhai, M. J. Efron, R. Iyer, M. C. Schatz, S. Sinha et G. E. Robinson, « Big Data : Astronomical or Genomical ? », *PLoS Biol*, t. 13, n° 7, e1002195, juil. 2015.

[15] P. Aldhous, « Managing the genome data deluge », *Science*, t. 262, n° 5133, p. 502-503, oct. 1993.

[16] T. Breur, « Statistical Power Analysis and the contemporary "crisis" in social sciences », *Journal of Marketing Analytics*, t. 4, n° 2, p. 61-65, juil. 2016, issn : 2050-3326. doi : 10.1057/s41270-016-0001-3. adresse : https://doi.org/10.1057/s41270-016-0001-3.

[17] R. Bellazzi, M. Diomidous, I. N. Sarkar, K. Takabayashi, A. Ziegler et A. T. McCray, « Data analysis and data mining : current issues in biomedical informatics », *Methods Inf Med*, t. 50, n° 6, p. 536-544, 2011.

[18] J. A. Blake et C. J. Bult, « Beyond the data deluge : data integration and bio-ontologies », *J Biomed Inform*, t. 39, n° 3, p. 314-320, juin 2006.

[19] T. Berners-Lee, J. Hendler et O. Lassila, « The Semantic Web : a new form of Web content that is meaningful to computers will unleash a revolution of new possibilities », *Scientific American*, t. 284, n° 5, p. 34-43, 2001, issn : 0036-8733. adresse : https://www.jstor.org/stable/26059207 (visité le 22/12/2018).

[20] H. De Jong, « Modeling and Simulation of Genetic Regulatory Systems : A Literature Review », *Journal of Computational Biology*, t. 9, n° 1, p. 67-103, 2002.

[21] D. A. Ouattara, W. Abou-Jaoudé et M. Kaufman, « From structure to dynamics : Frequency tuning in the p53-Mdm2 network. II : Differential and stochastic approaches », *Journal of Theoretical Biology*, t. 264, n° 4, p. 1177-1189, 2010. doi : 10.1016/j.jtbi.2010.03.031.

[22] S. Kauffman, « Metabolic stability and epigenesis in randomly constructed genetic nets », *Journal of Theoretical Biology*, t. 22, n° 3, p. 437-467, 1969, issn : 00225193. doi : 10.1016/0022-5193(69)90015-0.

[23] R. Thomas, « Regulatory networks seen as asynchronous automata : A logical description », *Journal of Theoretical Biology*, t. 153, n° 1, p. 1-23, 1991, issn : 00225193. doi : 10.1016/S0022-5193(05)80350-9.

[24] R. Thomas, D. Thieffry et M. Kaufman, « Dynamical behaviour of biological regulatory networks—I. Biological role of feedback loops and practical use of the concept of the loop-characteristic state », *Bulletin of Mathematical Biology*, t. 57, n° 2, p. 247-276, 1995, issn : 00928240. doi : 10.1016/0092-8240(94)00036-C.

[25] A. G. Gonzalez, A. Naldi, L. Sánchez, D. Thieffry et C. Chaouiya, « GINsim : A software suite for the qualitative modelling, simulation and analysis of regulatory networks », *BioSystems*, t. 84, n° 2, p. 91-100, 2006, issn : 03032647. doi : 10.1016/j.biosystems.2005.10.003.

[26] G. C. Tsokos, « Systemic lupus erythematosus », *N Engl J Med*, t. 365, n° 22, p. 2110-2121, déc. 2011.

[27] C. Bombardier, D. D. Gladman, M. B. Urowitz, D. Caron et C. H. Chang, « Derivation of the SLEDAI. A disease activity index for lupus patients. The Committee on Prognosis Studies in SLE », *Arthritis Rheum*, t. 35, n° 6, p. 630-640, juin 1992.

[28] E. M. Hay, P. A. Bacon, C. Gordon, D. A. Isenberg, P. Maddison, M. L. Snaith, D. P. Symmons, N. Viner et A. Zoma, « The BILAG index : a reliable and valid instrument for measuring clinical disease activity in systemic lupus erythematosus », *Q J Med*, t. 86, n° 7, p. 447-458, juil. 1993.

[29] M. Wahren-Herlenius et T. D ?rner, « Immunopathogenic mechanisms of systemic autoimmune disease », *Lancet*, t. 382, n° 9894, p. 819-831, août 2013.

[30] R. Felten, E. Dervovic, F. Chasset, J. E. Gottenberg, J. Sibilia, F. Scher et L. Arnaud, « The 2018 pipeline of targeted therapies under clinical development for Systemic Lupus Erythematosus : a systematic review of trials », *Autoimmun Rev*, t. 17, n° 8, p. 781-790, août 2018.

[31] L. Bennett, A. K. Palucka, E. Arce, V. Cantrell, J. Borvak, J. Banchereau et V. Pascual, « Interferon and granulopoiesis signatures in systemic lupus erythematosus blood », *J. Exp. Med.*, t. 197, n° 6, p. 711-723, mar. 2003.

[32] E. C. Baechler, F. M. Batliwalla, G. Karypis, P. M. Gaffney, W. A. Ortmann, K. J. Espe, K. B. Shark, W. J. Grande, K. M. Hughes, V. Kapur, P. K. Gregersen et T. W. Behrens, « Interferon-inducible gene expression signature in peripheral blood cells of patients with severe lupus », *Proc. Natl. Acad. Sci. U.S.A.*, t. 100, n° 5, p. 2610-2615, mar. 2003.

[33] R. Banchereau, S. Hong, B. Cantarel, N. Baldwin, J. Baisch, M. Edens, A.-M. Cepika, P. Acs, J. Turner, E. Anguiano, P. Vinod, S. Khan, G. Obermoser, D. Blankenship, E. Wakeland, L. Nassi, A. Gotte, M. Punaro, Y.-J. Liu, J. Banchereau, J. Rossello-Urgell, T. Wright et V. Pascual, « Personalized Immunomonitoring Uncovers Molecular Networks that Stratify Lupus Patients », *Cell*, t. 165, n° 3, p. 551-565, 2016, issn : 0092-8674. doi : https://doi.org/10.1016/j.cell.2016.03.008. adresse : http://www.sciencedirect.com/science/article/pii/S0092867416302641.

[34] G. Murphy et D. A. Isenberg, « Biologic therapies for systemic lupus erythematosus : where are we now ? », *Curr Opin Rheumatol*, t. 32, n° 6, p. 597-608, nov. 2020.

[35] Z. Touma et D. D. Gladman, « Current and future therapies for SLE : obstacles and recommendations for the development of novel treatments », *Lupus Sci Med*, t. 4, n° 1, e000239, 2017.

[36] S. V. Navarra, R. M. Guzm ?n, A. E. Gallacher, S. Hall, R. A. Levy, R. E. Jimenez, E. K. Li, M. Thomas, H. Y. Kim, M. G. Le ?n, C. Tanasescu, E. Nasonov, J. L. Lan, L. Pineda, Z. J. Zhong, W. Freimuth et M. A. Petri, « Efficacy and safety of belimumab in patients with active systemic lupus erythematosus : a randomised, placebo-controlled, phase 3 trial », *Lancet*, t. 377, n° 9767, p. 721-731, fév. 2011.

[37] S. Lal, A. E. McCart Reed, X. M. de Luca et P. T. Simpson, « Molecular signatures in breast cancer », *Methods*, t. 131, p. 135-146, déc. 2017.

[38] M. M. Kittleson et J. M. Hare, « Molecular signature analysis : using the myocardial transcriptome as a biomarker in cardiovascular disease », *Trends Cardiovasc. Med.*, t. 15, n° 4, p. 130-138, mai 2005.

[39] J. Sung, Y. Wang, S. Chandrasekaran, D. M. Witten et N. D. Price, « Molecular signatures from omics data : from chaos to consensus », *Biotechnol J*, t. 7, n° 8, p. 946-957, août 2012.

[40] H. Tang, S. Wang, G. Xiao, J. Schiller, V. Papadimitrakopoulou, J. Minna, I. I. Wistuba et Y. Xie, « Comprehensive evaluation of published gene expression prognostic signatures for biomarker-based lung cancer clinical studies », *Ann. Oncol.*, t. 28, n° 4, p. 733-740, avr. 2017.

[41] C. A. Borrebaeck, « Precision diagnostics : moving towards protein biomarker signatures of clinical utility in cancer », *Nat. Rev. Cancer*, t. 17, n° 3, p. 199-204, mar. 2017.

[42] J. Wang, Y. Zuo, Y. G. Man, I. Avital, A. Stojadinovic, M. Liu, X. Yang, R. S. Varghese, M. G. Tadesse et H. W. Ressom, « Pathway and network approaches for identification of cancer signature markers from omics data », *J Cancer*, t. 6, n° 1, p. 54-65, 2015.

[43] T. R. Samatov, V. V. Galatenko, A. Block, M. Y. Shkurnikov, A. G. Tonevitsky et U. Schumacher, « Novel biomarkers in cancer : The whole is greater than the sum of its parts », *Semin. Cancer Biol.*, t. 45, p. 50-57, août 2017.

[44] E. C. Baechler, F. M. Batliwalla, G. Karypis, P. M. Gaffney, W. A. Ortmann, K. J. Espe, K. B. Shark, W. J. Grande, K. M. Hughes, V. Kapur, P. K. Gregersen et T. W. Behrens, « Interferon-inducible gene expression signature in peripheral blood cells of patients with severe lupus », *Proc. Natl. Acad. Sci. U.S.A.*, t. 100, n° 5, p. 2610-2615, mar. 2003.

[45] T. Dierckx, L. Chiche, L. Daniel, B. Lauwerys, J. V. Weyenbergh et N. Jourde-Chiche, « Serum GlycA Level is Elevated in Active Systemic Lupus Erythematosus and Correlates to Disease Activity and Lupus Nephritis Severity », *J Clin Med*, t. 9, n° 4, mar. 2020.

[46] H. M. Kok, L. L. van den Hoogen, J. A. G. van Roon, E. J. M. Adriaansen, R. D. E. Fritsch-Stork, T. Q. Nguyen, R. Goldschmeding, T. R. D. J. Radstake et N. Bovenschen, « Systemic and local granzyme B levels are associated with disease activity, kidney damage and interferon signature in systemic lupus erythematosus », *Rheumatology (Oxford)*, t. 56, n° 12, p. 2129-2134, déc. 2017.

[47] E. Rouits, M. Boisdron-Celle, A. Dumont, O. Gu ?rin, A. Morel et E. Gamelin, « Relevance of different UGT1A1 polymorphisms in irinotecan-induced toxicity : a molecular and clinical study of 75 patients », *Clin. Cancer Res.*, t. 10, n° 15, p. 5151-5159, août 2004.

[48] M. Bertolo, S. Baumgart, P. Durek, A. Peddinghaus, H. Mei, T. Rose, P. Enghard et A. Gr ?tzkau, « Deep Phenotyping of Urinary Leukocytes by Mass Cytometry Reveals a Leukocyte Signature for Early and Non-Invasive Prediction of Response to Treatment in Active Lupus Nephritis », *Front Immunol*, t. 11, p. 256, 2020.

[49] F. Drakopanagiotakis, L. Wujak, M. Wygrecka et P. Markart, « Biomarkers in idiopathic pulmonary fibrosis », *Matrix Biol.*, t. 68-69, p. 404-421, août 2018.

[50] M. Ferro, P. Ungaro, A. Cimmino, G. Lucarelli, G. M. Busetto, F. Cantiello, R. Damiano et D. Terracciano, « Epigenetic Signature : A New Player as Predictor of Clinically Significant Prostate Cancer (PCa) in Patients on Active Surveillance (AS) », *Int J Mol Sci*, t. 18, n° 6, mai 2017.

[51] D. Ghosh et L. M. Poisson, « "Omics" data and levels of evidence for biomarker discovery », *Genomics*, t. 93, n° 1, p. 13-16, jan. 2009.

[52] B. Freidlin et R. Simon, « Adaptive signature design : an adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients », *Clin. Cancer Res.*, t. 11, n° 21, p. 7872-7878, nov. 2005.

[53] C. Guo, X. Lin, J. Yin, X. Xie, J. Li, X. Meng, J. Wu, L. Huang, Z. Huang, G. Yang, H. Zhou et X. Chen, « Pharmacogenomics signature : A novel strategy on the individual differences in drug response », *Cancer Lett.*, t. 420, p. 190-194, avr. 2018.

[54] L. Maiuri, V. Raia et G. Kroemer, « Strategies for the etiological therapy of cystic fibrosis », *Cell Death Differ.*, t. 24, n° 11, p. 1825-1844, nov. 2017.

[55] E. R. Fearon, « Molecular genetics of colorectal cancer », *Annu Rev Pathol*, t. 6, p. 479-507, 2011.

[56] X. Liu, Y. Wang, H. Ji, K. Aihara et L. Chen, « Personalized characterization of diseases using sample-specific networks », *Nucleic Acids Res.*, t. 44, n° 22, e164, déc. 2016.

[57] K. Zhu, C. Pian, Q. Xiang, X. Liu et Y. Chen, « Personalized analysis of breast cancer using sample-specific networks », *PeerJ*, t. 8, e9161, 2020.

[58] F. Hu, Q. Wang, Z. Yang, Z. Zhang et X. Liu, « Network-based identification of biomarkers for colon adenocarcinoma », *BMC Cancer*, t. 20, n° 1, p. 668, juil. 2020.

[59] K. L. Buschur, M. Chikina et P. V. Benos, « Causal network perturbations for instance-specific analysis of single cell and disease samples », *Bioinformatics*, t. 36, n° 8, p. 2515-2521, avr. 2020.

[60] J. Béal, A. Montagud, P. Traynard, E. Barillot et L. Calzone, « Personalization of Logical Models With Multi-Omics Data Allows Clinical Stratification of Patients », *Frontiers in Physiology*, t. 9, p. 1965, 2019, issn : 1664-042X. doi : 10.3389/fphys.2018.01965. adresse : https://www.frontiersin.org/article/10.3389/fphys.2018.01965.

[61] F. Eduati, V. Dold ?n-Martelli, B. Klinger, T. Cokelaer, A. Sieber, F. Kogera, M. Dorel, M. J. Garnett, N. Bl ?thgen et J. Saez-Rodriguez, « Drug Resistance Mechanisms in Colorectal Cancer Dissected with Cell Type-Specific Dynamic Logic Models », *Cancer Res.*, t. 77, n° 12, p. 3364-3375, juin 2017.

[62] D. A. Barbie, P. Tamayo, J. S. Boehm, S. Y. Kim, S. E. Moody, I. F. Dunn, A. C. Schinzel, P. Sandy, E. Meylan, C. Scholl, S. Fr ?hling, E. M. Chan, M. L. Sos, K. Michel, C. Mermel, S. J. Silver, B. A. Weir, J. H. Reiling, Q. Sheng, P. B. Gupta, R. C. Wadlow, H. Le, S. Hoersch, B. S. Wittner, S. Ramaswamy, D. M. Livingston, D. M. Sabatini, M. Meyerson, R. K. Thomas, E. S. Lander, J. P. Mesirov, D. E. Root, D. G. Gilliland, T. Jacks et W. C. Hahn, « Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1 », *Nature*, t. 462, n° 7269, p. 108-112, nov. 2009.

[63] Y. Drier, M. Sheffer et E. Domany, « Pathway-based personalized analysis of cancer », *Proc. Natl. Acad. Sci. U.S.A.*, t. 110, n° 16, p. 6388-6393, avr. 2013.

[64] I. Kuperstein, E. Bonnet, H. A. Nguyen, D. Cohen, E. Viara, L. Grieco, S. Fourquet, L. Calzone, C. Russo, M. Kondratova, M. Dutreix, E. Barillot et A. Zinovyev, « Atlas of Cancer Signalling Network : a systems biology resource for integrative analysis of cancer data with Google Maps », *Oncogenesis*, t. 4, e160, juil. 2015.

[65] E. R. Holzinger, S. M. Dudek, A. T. Frase, S. A. Pendergrass et M. D. Ritchie, « ATHENA : the analysis tool for heritable and environmental network associations », *Bioinformatics*, t. 30, n° 5, p. 698-705, mar. 2014.

[66] B. L. Fridley, S. Lund, G. D. Jenkins et L. Wang, « A Bayesian integrative genomic model for pathway analysis of complex traits », *Genet Epidemiol*, t. 36, n° 4, p. 352-359, mai 2012.

[67] P. K. Mankoo, R. Shen, N. Schultz, D. A. Levine et C. Sander, « Time to recurrence and survival in serous ovarian tumors predicted from integrated genomic profiles », *PLoS One*, t. 6, n° 11, e24709, 2011.

[68] D. Kim, J. G. Joung, K. A. Sohn, H. Shin, Y. R. Park, M. D. Ritchie et J. H. Kim, « Knowledge boosting : a graph-based integration approach with multi-omics data and genomic knowledge for cancer clinical outcome prediction », *J Am Med Inform Assoc*, t. 22, n° 1, p. 109-120, jan. 2015.

[69] G. R. Lanckriet, T. De Bie, N. Cristianini, M. I. Jordan et W. S. Noble, « A statistical framework for genomic data fusion », *Bioinformatics*, t. 20, n° 16, p. 2626-2635, nov. 2004.

[70] H. Sharifi-Noghabi, O. Zolotareva, C. C. Collins et M. Ester, « MOLI : multi-omics late integration with deep neural networks for drug response prediction », *Bioinformatics*, t. 35, n° 14, p. i501-i509, juil. 2019.

[71] A. Singh, C. P. Shannon, B. Gautier, F. Rohart, M. Vacher, S. J. Tebbutt et K. A. L ? Cao, « DIABLO : an integrative approach for identifying key molecular drivers from multi-omics assays », *Bioinformatics*, t. 35, n° 17, p. 3055-3062, sept. 2019.

[72] A. Dugourd et J. Saez-Rodriguez, « Footprint-based functional analysis of multiomic data », *Curr Opin Syst Biol*, t. 15, p. 82-90, juin 2019.

[73] M. D. Ritchie, E. R. Holzinger, R. Li, S. A. Pendergrass et D. Kim, « Methods of integrating data to uncover genotype-phenotype interactions », *Nat. Rev. Genet.*, t. 16, n° 2, p. 85-97, fév. 2015.

[74] G. Tini, L. Marchetti, C. Priami et M. P. Scott-Boyer, « Multi-omics integration-a comparison of unsupervised clustering methodologies », *Brief. Bioinformatics*, t. 20, n° 4, p. 1269-1279, juil. 2019.

[75] K. K. Yan, H. Zhao et H. Pang, « A comparison of graph- and kernel-based -omics data integration algorithms for classifying complex traits », *BMC Bioinformatics*, t. 18, n° 1, p. 539, déc. 2017.

[76] R. Haring et H. Wallaschofski, « Diving through the "-omics" : the case for deep phenotyping and systems epidemiology », *OMICS*, t. 16, n° 5, p. 231-234, mai 2012.

[77] N. Ishii, K. Nakahigashi, T. Baba, M. Robert, T. Soga, A. Kanai, T. Hirasawa, M. Naba, K. Hirai, A. Hoque, P. Y. Ho, Y. Kakazu, K. Sugawara, S. Igarashi, S. Harada, T. Masuda, N. Sugiyama, T. Togashi, M. Hasegawa, Y. Takai, K. Yugi, K. Arakawa, N. Iwata, Y. Toya, Y. Nakayama, T. Nishioka, K. Shimizu, H. Mori et M. Tomita, « Multiple high-throughput analyses monitor the response of E. coli to perturbations », *Science*, t. 316, n° 5824, p. 593-597, avr. 2007.

[78] B. L. Parker, A. C. Calkin, M. M. Seldin, M. F. Keating, E. J. Tarling, P. Yang, S. C. Moody, Y. Liu, E. J. Zerenturk, E. J. Needham, M. L. Miller, B. L. Clifford, P. Morand, M. J. Watt, R. C. R. Meex, K. Y. Peng, R. Lee, K. Jayawardana, C. Pan, N. A. Mellett, J. M. Weir, R. Lazarus, A. J. Lusis, P. J. Meikle, D. E. James, T. Q. de Aguiar Vallim et B. G. Drew, « An integrative systems genetic analysis of mammalian lipid metabolism », *Nature*, t. 567, n° 7747, p. 187-193, mar. 2019.

[79] J. Escoubet, M. Kenigsberg, M. Derock, V. Yaligara, M. D. Bock, S. Roche, F. Massey, H. de Foucauld, C. Bettembourg, A. Olivier, A. Berthemy, J. Capdevielle, R. Legoux, E. Perret, A. Buzy, P. Chardenot, V. Destelle, A. Leroy, C. Cahours, S. Teixeira, P. Juvet, P. Gauthier, M. Leguet, L. Rocheteau-Beaujouan, M. A. Chatoux, W. Deshayes, M. Clement, M. Kabiri, C. Orsini, V. Mikol, M. Didier et J. C. Guillemot, « ABHD11, a new diacylglycerol lipase involved in weight gain regulation », *PLoS One*, t. 15, n° 6, e0234780, 2020.

[80] Q. Wu, J. V. Li, F. Seyfried, C. W. le Roux, H. Ashrafian, T. Athanasiou, W. Fenske, A. Darzi, J. K. Nicholson, E. Holmes et N. J. Gooderham, « Metabolic phenotype-microRNA data fusion analysis of the systemic consequences of Roux-en-Y gastric bypass surgery », *Int J Obes (Lond)*, t. 39, n° 7, p. 1126-1134, juil. 2015.

[81] K. Yugi, H. Kubota, A. Hatano et S. Kuroda, « Trans-Omics : How To Reconstruct Biochemical Networks Across Multiple 'Omic' Layers », *Trends Biotechnol*, t. 34, n° 4, p. 276-290, avr. 2016.

[82] H. F. Deus, D. F. Veiga, P. R. Freire, J. N. Weinstein, G. B. Mills et J. S. Almeida, « Exposing the cancer genome atlas as a SPARQL endpoint », *J Biomed Inform*, t. 43, n° 6, p. 998-1008, déc. 2010.

[83] M. Saleem, S. S. Padmanabhuni, A. C. Ngomo, A. Iqbal, J. S. Almeida, S. 12ker et H. F. Deus, « TopFed : TCGA tailored federated query processing and linking to LOD », *J Biomed Semantics*, t. 5, p. 47, 2014.

[84] N. S. Miyoshi, D. G. Pinheiro, W. A. Silva et J. C. Felipe, « Computational framework to support integration of biomolecular and clinical data within a translational approach », *BMC Bioinformatics*, t. 14, p. 180, juin 2013.

[85] M. E. Holford, J. P. McCusker, K. H. Cheung et M. Krauthammer, « A semantic web framework to integrate cancer omics data with biological knowledge », *BMC Bioinformatics*, t. 13 Suppl 1, S10, jan. 2012.

[86] N. I. Panousis, G. K. Bertsias, H. Ongen, I. Gergianaki, M. G. Tektonidou, M. Trachana, L. Romano-Palumbo, D. Bielser, C. Howald, C. Pamfil, A. Fanouriakis, D. Kosmara, A. Repa, P. Sidiropoulos, E. T. Dermitzakis et D. T. Boumpas, « Combined genetic and transcriptome analysis of patients with SLE : distinct, targetable signatures for susceptibility and severity », *Annals of the Rheumatic Diseases*, t. 78, n° 8, p. 1079-1089, 2019, issn : 0003-4967.

[87] S. Jung, A. Hartmann et A. Del Sol, « RefBool : a reference-based algorithm for discretizing gene expression data », *Bioinformatics*, t. 33, n° 13, p. 1953-1962, août 2017.

[88] M. E. Goddard, K. E. Kemper, I. M. MacLeod, A. J. Chamberlain et B. J. Hayes, « Genetics of complex traits : prediction of phenotype, identification of causal polymorphisms and genetic architecture », *Proc. Biol. Sci.*, t. 283, n° 1835, juil. 2016.

[89] M. Morley, C. M. Molony, T. M. Weber, J. L. Devlin, K. G. Ewens, R. S. Spielman et V. G. Cheung, « Genetic analysis of genome-wide variation in human gene expression », *Nature*, t. 430, n° 7001, p. 743-747, août 2004.

[90] G. consortium, « Genetic effects on gene expression across human tissues », *Nature*, t. 550, n° 7675, p. 204-213, oct. 2017.

[91] S. Lê, J. Josse et F. Husson, « FactoMineR : A Package for Multivariate Analysis », *Journal of Statistical Software*, t. 25, n° 1, p. 1-18, 2008. doi : 10.18637/jss.v025.i01.

[92] G. Consortium, « Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis : multitissue gene regulation in humans », *Science*, t. 348, n° 6235, p. 648-660, mai 2015.

[93] Z. Zhu, F. Zhang, H. Hu, A. Bakshi, M. R. Robinson, J. E. Powell, G. W. Montgomery, M. E. Goddard, N. R. Wray, P. M. Visscher et J. Yang, « Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets », *Nat. Genet.*, t. 48, n° 5, p. 481-487, mai 2016.

[94] E. Porcu, S. R ?eger, K. Lepik, F. A. Santoni, A. Reymond, Z. Kutalik, M. Agbessi, H. Ahsan, I. Alves, A. Andiappan, W. Arindrarto, P. Awadalla, A. Battle, F. Beutner, M. 01 Bonder, D. Boomsma, M. Christiansen, A. Claringbould, P. Deelen, T. Esko, M. J. Fav ?, L. Franke, T. Frayling, S. A. Gharib, G. Gibson, B. T. Heijmans, G. Hemani, R. 01sen, M. K ?h ?nen, A. Kalnapenkis, S. Kasela, J. Kettunen, Y. Kim, H. Kirsten, P. Kovacs, K. Krohn, J. Kronberg-Guzman, V. Kukushkina, B. Lee, T. Lehtim ?ki, M. Loeffler, U. M. Marigorta, H. Mei, L. Milani, G. W. Montgomery, M. M ?ller-Nurasyid, M. Nauck, M. Nivard, B. Penninx, M. Perola, N. Pervjakova, B. L. Pierce, J. Powell, H. Prokisch, B. M. Psaty, O. T. Raitakari, S. Ripatti, O. Rotzschke, A. Saha, M. Scholz, K. Schramm, I. Sepp ?l ?, E. P. Slagboom, C. D. A. Stehouwer, M. Stumvoll, P. Sullivan, P. A. C. 't Hoen, A. Teumer, J. Thiery, L. Tong, A. T ?njes, J. van Dongen, M. van Iterson, J. van Meurs, J. H. Veldink, J. Verlouw, P. M. Visscher, U. V ?lker, U. V ?sa, H. J. Westra, C. Wijmenga, H. Yaghootkar, J. Yang, B. Zeng, F. Zhang, W. Arindrarto, M. Beekman, D. I. Boomsma, J. Bot, J. Deelen, P. Deelen, L. Franke, B. T. Heijmans, P. A. C. 't Hoen, B. A. Hofman, J. J. Hottenga, A. Isaacs, M. J. Bonder, P. M. Jhamai, R. 01sen, S. M. Kielbasa, N. Lakenberg, R. Luijk, H. Mei, M. Moed, I. Nooren, R. Pool, C. G. Schalkwijk, P. E. Slagboom, C. D. A. Stehouwer, H. E. D. Suchiman, M. A. Swertz, E. F. Tigchelaar, A. G. Uitterlinden, L. H. van den Berg, R. van der Breggen, C. J. H. van der Kallen, F. van Dijk, J. van Dongen, C. M. van Duijn, M. van Galen, M. M. J. van Greevenbroek, D. van Heemst, M. van Iterson, J. van Meurs, J. van Rooij, P. Van't Hof, E. W. van Zwet, M. Vermaat, J. H. Veldink, M. Verbiest, M. Verkerk, C. Wijmenga, D. V. Zhernakova et S. Zhernakova, « Mendelian randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits », *Nat Commun*, t. 10, n° 1, p. 3300, juil. 2019.

[95] A. Polynikis, S. J. Hogan et M. di Bernardo, « Comparing different ODE modelling approaches for gene regulatory networks », *J. Theor. Biol.*, t. 261, n° 4, p. 511-530, 2009.

[96] G. Yagil, « Quantitative aspects of protein induction », in *Current topics in Cell regulation*, B. Horecker et E. Stadtman, éd., Academic Press, 1975, p. 183-237.

[97] R. de Sousa Abreu, L. O. Penalva, E. M. Marcotte et C. Vogel, « Global signatures of protein and mRNA expression levels », *Mol Biosyst*, t. 5, n° 12, p. 1512-1526, 2009.

[98] W. Abou-Jaoudé, D. A. Ouattara et M. Kaufman, « From structure to dynamics : Frequency tuning in the p53–Mdm2 network : I. Logical approach », *Journal of Theoretical Biology*, t. 258, n° 4, p. 561-577, 2009, issn : 0022-5193. doi : https://doi.org/10.1016/j.jtbi.2009.02.005.

[99] A. Fauré, A. Naldi, C. Chaouiya et D. Thieffry, « Dynamical analysis of a generic Boolean model for the control of the mammalian cell cycle », *Bioinformatics*, t. 22, n° 14, e124-e131, 2006. doi : 10.1093/bioinformatics/btl210. eprint : /oup/backfile/content_public/journal/bioinformatics/22/14/10.1093/bioinformatics/btl210/2/btl210.pdf.

[100] E. Remy, S. Rebouissou, C. Chaouiya, A. Zinovyev, F. Radvanyi et L. Calzone, « A Modeling Approach to Explain Mutually Exclusive and Co-Occurring Genetic Alterations in Bladder Tumorigenesis », *Cancer research*, t. 75, n° 19, p. 4042-52, 2015, issn : 1538-7445. doi : 10.1158/0008-5472.CAN-15-0602.

[101] A. Naldi, J. Carneiro, C. Chaouiya et D. Thieffry, « Diversity and plasticity of Th cell types predicted from regulatory network modelling », *PLoS Computational Biology*, t. 6, n° 9, 2010, issn : 1553734X. doi : 10.1371/journal.pcbi.1000912.

[102] W. Abou-Jaoudé, P. T. Monteiro, A. Naldi, M. Grandclaudon, V. Soumelis, C. Chaouiya et D. Thieffry, « Model checking to assess T-helper cell plasticity. », *Frontiers in bioengineering and biotechnology*, t. 2, p. 86, 2014. doi : 10.3389/fbioe.2014.00086.

[103] A. Liberzon, C. Birger, H. Thorvaldsdottir, M. Ghandi, J. P. Mesirov et P. Tamayo, « The Molecular Signatures Database (MSigDB) hallmark gene set collection », *Cell Syst*, t. 1, n° 6, p. 417-425, 2015.

[104] R. Samaga, A. Von Kamp et S. Klamt, « Computing combinatorial intervention strategies and failure modes in signaling networks », *J. Comput. Biol.*, t. 17, n° 1, p. 39-53, 2010.

[105] N. Levy, A. Naldi, C. Hernandez, G. Stoll, D. Thieffry, A. Zinovyev, L. Calzone et L. Paulevé, « Prediction of Mutations to Control Pathways Enabling Tumour Cell Invasion with the CoLoMoTo Interactive Notebook (Tutorial) », *Frontiers in Physiology*, t. 9, p. 787, 2018. doi : 10.3389/fphys.2018.00787.

[106] M. Folschette, L. Paulevé, M. Magnin et O. Roux, « Sufficient conditions for reachability in automata networks with priorities », *Theoretical Computer Science*, t. 608, Part 1, From Computer Science to Biology and Back, p. 66-83, 2015, issn : 0304-3975. doi : 10.1016/j.tcs.2015.08.040.

[107] D. Fang et J. Zhu, « Dynamic balance between master transcription factors determines the fates and functions of CD4 T cell and innate lymphoid cell subsets. », *The Journal of experimental medicine*, t. 214, n° 7, p. 1861-1876, 2017.

[108] S. Yepes, M. M. Torres et R. E. Andrade, « Clustering of Expression Data in Chronic Lymphocytic Leukemia Reveals New Molecular Subdivisions. », *PloS one*, t. 10, n° 9, e0137132, 2015, issn : 1932-6203. doi : 10.1371/journal.pone.0137132.

[109] B. Ganter et R. Wille, *Formal concept analysis : mathematical foundations*. Springer, 1999, p. 284, isbn : 3540627715.

[110] E. Remy, P. Ruet, L. Mendoza, D. Thieffry et C. Chaouiya, « From Logical Regulatory Graphs to Standard Petri Nets : Dynamical Roles and Functionality of Feedback Circuits », in *Transactions on Computational Systems Biology VII*, C. Priami, A. Ingólfsdóttir, B. Mishra et H. Riis Nielson, éd., Berlin, Heidelberg : Springer Berlin Heidelberg, 2006, p. 56-72, isbn : 978-3-540-48839-2. doi : 10.1007/11905455_3.

[111] M. Alam, T. N. N. Le et A. Napoli, « LatViz : A New Practical Tool for Performing Interactive Exploration over Concept Lattices », in *Proceedings of the Thirteenth International Conference on Concept Lattices and Their Applications, Moscow, Russia, July 18-22, 2016.*, 2016, p. 9-20.

[112] B. Ganter, G. Stumme et R. Wille, *Formal concept analysis : foundations and applications*. Springer, 2005, p. 348, isbn : 354031881X.

[113] A. Lihonosova et A. Kaminskaya, « Using Formal Concept Analysis for Finding the Closest Relatives among a Group of Organisms », English, *Procedia Computer Science*, t. 31, n° Complete, p. 860-868, 2014. doi : 10.1016/j.procs.2014.05.337.

[114] L. Bourneuf et J. Nicolas, « FCA in a Logical Programming Setting for Visualization-Oriented Graph Compression », in *Formal Concept Analysis*, K. Bertet, D. Borchmann, P. Cellier et S. Ferré, éd., Cham : Springer International Publishing, 2017, p. 89-105, isbn : 978-3-319-59271-8.

[115] V. Wucher, D. Tagu et J. Nicolas, « Edge Selection in a Noisy Graph by Concept Analysis : Application to a Genomic Network », in *Data Science, Learning by Latent Structures, and Knowledge Discovery*, B. Lausen, S. Krolak-Schwerdt et M. Böhmer, éd., Berlin, Heidelberg : Springer Berlin Heidelberg, 2015, p. 353-364, isbn : 978-3-662-44983-7.

[116] J. Sheng, Q. Chen, I. Soncin, S. L. Ng, K. Karjalainen et C. Ruedl, « A Discrete Subset of Monocyte-Derived Cells among Typical Conventional Type 2 Dendritic Cells Can Efficiently Cross-Present », *Cell Reports*, t. 21, n° 5, p. 1203-1214, 2017.

[117] E. Mitsi, R. Kamng'ona, J. Rylance, C. Solórzano, J. Jesus Reiné, H. C. Mwandumba, D. M. Ferreira et K. C. Jambo, « Human alveolar macrophages predominately express combined classical M1 and M2 surface markers in steady state. », *Respiratory research*, t. 19, n° 1, p. 66, 2018.

[118] J.-P. Barthélémy et A. Guénoche, *Trees and proximity representations*. John Wiley & Sons, 1991.

[119] T. R. Mosmann et R. L. Coffman, « TH1 and TH2 Cells : Different Patterns of Lymphokine Secretion Lead to Different Functional Properties », *Annual Review of Immunology*, t. 7, n° 1, p. 145-173, 1989, issn : 0732-0582. doi : 10.1146/annurev.iy.07.040189.001045.

[120] D. A. A. Vignali, L. W. Collison et C. J. Workman, « How regulatory T cells work. », *Nature reviews. Immunology*, t. 8, n° 7, p. 523-32, 2008, issn : 1474-1741. doi : 10.1038/nri2343.

[121] E. A. Ivanova et A. N. Orekhov, « T Helper Lymphocyte Subsets and Plasticity in Autoimmunity and Cancer : An Overview. », *BioMed research international*, t. 2015, p. 327 470, 2015, issn : 2314-6141. doi : 10.1155/2015/327470.

[122] T. Caza et S. Landas, « Functional and Phenotypic Plasticity of CD4(+) T Cell Subsets. », *BioMed research international*, t. 2015, p. 521 957, 2015, issn : 2314-6141. doi : 10.1155/2015/521957.

[123] V. T. Thieu, Q. Yu, H.-C. Chang, N. Yeh, E. T. Nguyen, S. Sehra et M. H. Kaplan, « Signal transducer and activator of transcription 4 is required for the transcription factor T-bet to promote T helper 1 cell-fate determination. », *Immunity*, t. 29, n° 5, p. 679-90, 2008.

[124] Y. Wang, M. A. Su et Y. Y. Wan, « An essential role of the transcription factor GATA-3 for the function of regulatory T cells », *Immunity*, t. 35, n° 3, p. 337-348, sept. 2011.

UNIVERSITÉ DE RENNES 1

Titre : Identification de signature causale pathologique par intégration de données multi-omiques

Mot clés : signature causale, données omiques, intégration, réseau biologique

Résumé : Le lupus systémique erythémateux est un exemple de maladie complexe, hétérogène et multi-factorielle. L'identification de signature pouvant expliquer la cause d'une maladie est un enjeu important pour la stratification des patients. De plus, les analyses statistiques classiques s'appliquent difficilement quand les populations d'intérêt sont hétérogènes et ne permettent pas de mettre en évidence la cause. Cette thèse présente donc deux méthodes permettant de répondre à cette problématique. Tout d'abord, un modèle transomique est décrit pour structurer l'ensemble des données omiques en utilisant le Web sémantique (RDF). Son alimentation repose sur une analyse à l'échelle du patient. L'interrogation de ce modèle sous forme d'une requête SPARQL a permis l'identification d'expression Individually-Consistent Trait Loci (eICTLs). Il s'agit d'une association par raisonnement d'un couple SNP-gène pour lequel la présence d'un SNP influence la variation d'expression du gène. Ces éléments ont permis de réduire la dimensionalité des données omiques et présentent un apport plus informatif que les données de génomique. Cette première méthode se base uniquement sur l'utilisation des données omiques. Ensuite, la deuxième méthode repose sur la dépendance entre les régulations existante dans les réseaux biologiques. En combinant la dynamique des systèmes biologiques et l'analyse par concept formel, les états stables générés sont automatiquement classés. Cette classification a permis d'enrichir des signatures biologiques, caractéristique de phénotype. De plus, de nouveaux phénotypes hybrides ont été identifiés.

Title: Identification of causal pathologic signature by multi-omic data integration

Keywords: causal signature, omics data, integration, biological network

Abstract: Systematic erythematosus lupus is an example of a complex, heterogeneous and multifactorial disease. The identification of signature that can explain the cause of a disease remains an important challenge for the stratification of patients. Classic statistical analysis can hardly be applied when population of interest are heterogeneous and they do not highlight the cause. This thesis presents two methods that answer those issues. First, a transomic model is described in order to structure all the omic data, using semantic Web (RDF). Its supplying is based on a patient-centric approach. SPARQL query interrogates this model and allow the identification of expression Individually-Consistent Trait Loci (eICTLs). It a reasoning association between a SNP and a gene whose the presence of the SNP impact the variation of its gene expression. Those elements provide a reduction of omics data dimension and show a more informative contribution than genomic data. This first method are omics data-driven. Then, the second method is based on the existing regulation dependancies in biological networks. By combining the dynamic of biological system with the formal concept analysis, the generated stable states are automatically classified. This classification enables the enrichment of biological signature, which caracterised a phenotype. Moreover, new hybrid phenotype is identified.