

Les maladies complexes ou multifactorielles sont définies comme des maladies dont l'étiologie et la progression sont déterminées par un certain nombre de facteurs génétiques, environnementaux et épigénétiques. On peut citer à cet égard les maladies auto-immunes comme la polyarthrite rhumatoïde (PR), les maladies neurologiques comme la maladie d'Alzheimer, la maladie de Parkinson et la sclérose en plaques, les maladies respiratoires comme l'asthme, les maladies du tissu conjonctif comme la sclérodermie, les maladies rénales, et bien d'autres encore. Bien que les chercheurs dans le domaine de la biomédecine aient fait des progrès significatifs contre diverses autres maladies mortelles, ils sont encore confrontés à des défis pour comprendre les mécanismes fondamentaux sous-jacents l'initiation et de la progression de maladies complexes. La raison en est que la méthodologie de la biologie expérimentale classique est principalement basée sur l'étude de gènes et de protéines individuels, en considérant l'organisme comme simple et linéaire. Cependant, les maladies complexes impliquent de nombreuses interactions complexes entre les différentes entités biologiques, ce qui entraîne certaines conséquences pathologiques critiques. Dans ce projet, nous avons essayé de comprendre le mécanisme sous-jacent l'initiation d'une maladie complexe courante, la polyarthrite rhumatoïde (PR). La PR touche environ 0.5-1.0% de la population adulte dans la plupart des pays développés. La prévalence mondiale de la PR serait un peu plus faible (environ 0.24%), avec une prévalence trois fois plus élevée chez les femmes que chez les hommes et environ 0.5 à 1% chez les individus blancs selon une étude basée sur les pays occidentaux.

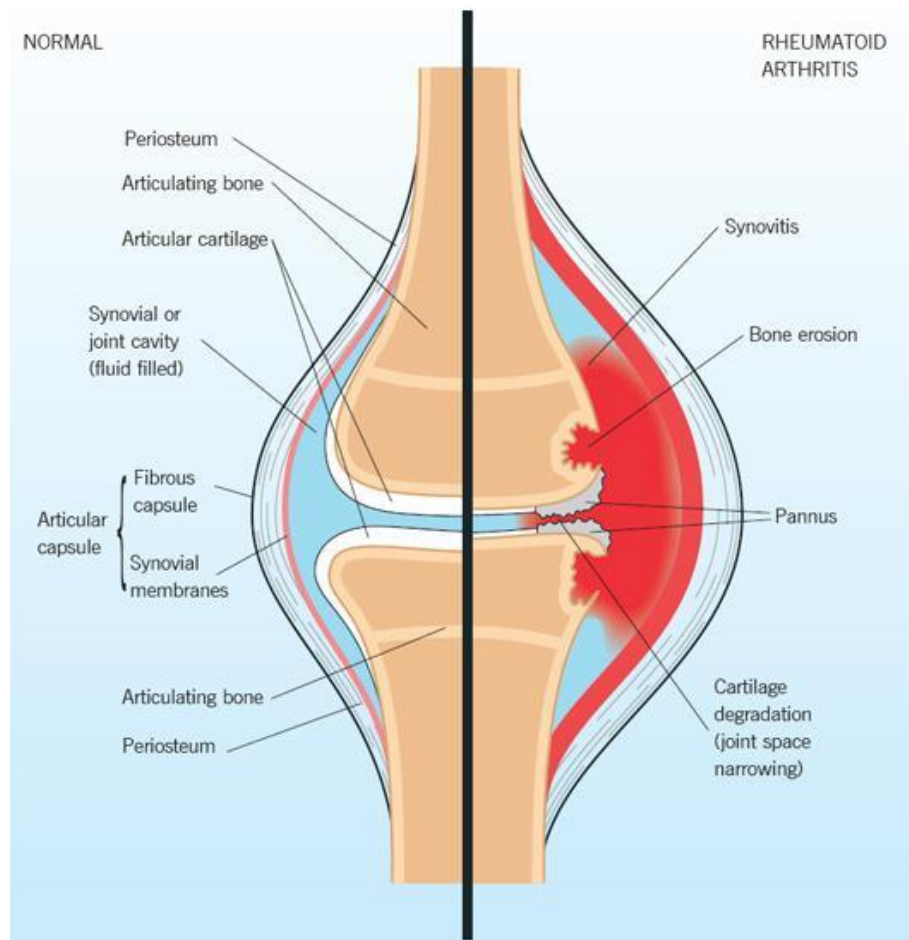


Figure 1: The pathology of RA (figure taken from (Clancy & Hasthorpe, 2011)).

La PR est une maladie auto-immune inflammatoire chronique dont la pathogenèse implique une interaction complexe entre des facteurs génétiques, épigénétiques et environnementaux, entraînant une cascade de réactions immunitaires. La PR affecte principalement le cartilage et les articulations synoviales des mains, des poignets et des pieds, des épaules, des coudes, des genoux et des chevilles. La PR est caractérisée par une inflammation et une hyperplasie synoviale, la production d'auto-anticorps (facteur rhumatoïde [FR] et anticorps protéiques anti-citrullinés [ACPA]), la destruction du cartilage et des os, et des caractéristiques systémiques, notamment des troubles cardiovasculaires, pulmonaires, psychologiques et squelettiques. La complexité des multiples voies de signalisation qui interagissent les unes avec les autres, entraînant l'expression de diverses molécules pro-inflammatoires comme les cytokines, chimiokines et métalloprotéinases matricielles (MMP), rend l'étude de l'étiologie de la maladie quelque peu déroutante. Les processus du système immunitaire inné et adaptatif jouent un rôle majeur dans la propagation des signaux qui entraînent l'activation de multiples cellules immunitaires comme les macrophages, les neutrophiles, les cellules T et les cellules B, ainsi que l'activation de cellules articulaires comme les fibroblastes synoviaux. L'infiltration et l'accumulation de cellules inflammatoires comme les macrophages, les lymphocytes (lymphocytes T, lymphocytes B) et les neutrophiles entraînent une hyperplasie de la membrane synoviale et, par la suite, la formation de pannus. En plus des diverses cellules immunitaires étudiées pour l'intervention thérapeutique dans la maladie, les fibroblastes synoviaux (FS) sont apparus comme une cible thérapeutique supplémentaire. De nombreuses études se concentrent sur les FS pour mieux comprendre les interactions complexes avec les cellules immunitaires et comment la multiplication des FS peut éventuellement induire la destruction du cartilage et l'érosion osseuse dans les articulations. En général, dans les tissus sains, les fibroblastes synoviaux fournissent à la cavité articulaire et au cartilage adjacent des protéines plasmatiques et des molécules lubrifiantes telles que l'acide hyaluronique. Les FS produisent également des protéines de matrice telles que le collagène et l'hyaluronane, une variété d'enzymes dégradant la matrice, comme les métalloprotéases de matrice (MMPS) qui contribuent au remodelage de la matrice.

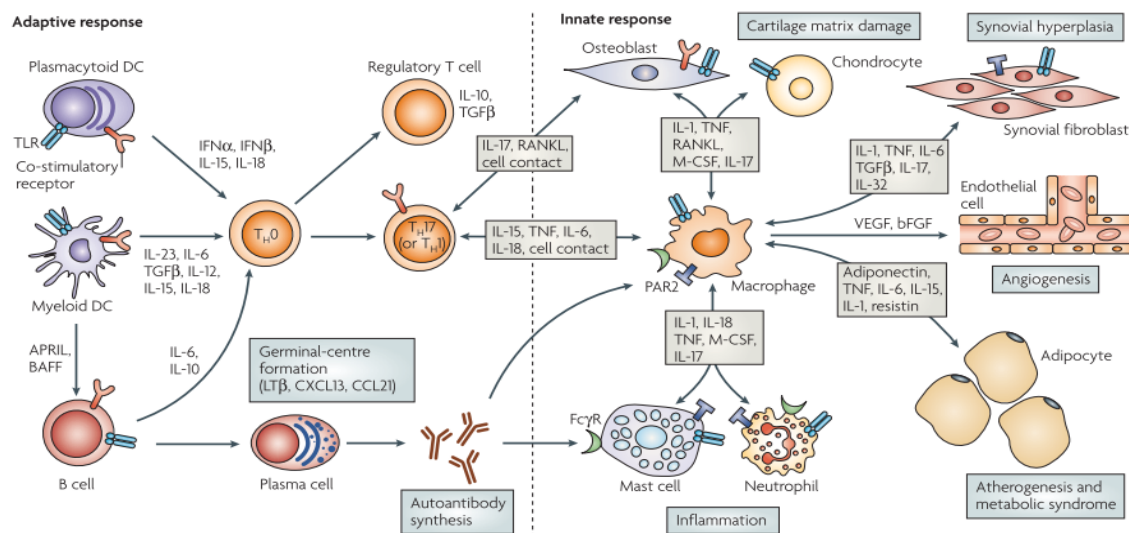


Figure 4: Cytokine mediated synovial regulation and interaction between various cells of both innate and adaptive immune systems during the progression of RA (taken from (McInnes & Schett, 2007))

Le diagnostic de la PR est principalement basé sur les manifestations cliniques et les biomarqueurs sériques tels que le FR et l'ACPA, ce dernier étant plus spécifique à la maladie. La positivité de l'ACPA est maintenant largement utilisée pour diagnostiquer la PR dans la pratique clinique en raison de sa grande spécificité (>97%). Un examen médical typique comprend une recherche de gonflement, de marasme et de sensibilité avec atrophie des muscles à proximité des articulations concernées. Les niveaux de protéine C réactive et la vitesse de sédimentation des érythrocytes sont souvent augmentés en cas de PR active, et ces réactifs en phase aiguë font partie des nouveaux critères de classification de la PR. L'hérédité de la PR est d'environ 60% avec des facteurs génétiques comme les antigènes majeurs d'histocompatibilité de classe II/antigènes de leucocytes humains (HLA-DR) et les gènes non-HLA associés à sa pathogénèse. Les allèles épitopes partagés qui sont le plus significativement associés à la PR sont HLA-DRB1*01 et HLA-DRB1*04. Les gènes non-HLA les plus pertinents avec des polymorphismes de nucléotides uniques (SNP) associés à la PR sont PTPN22, IL23R, TRAF1, CTLA4, IRF5, STAT4, CCR6 et PADI4.

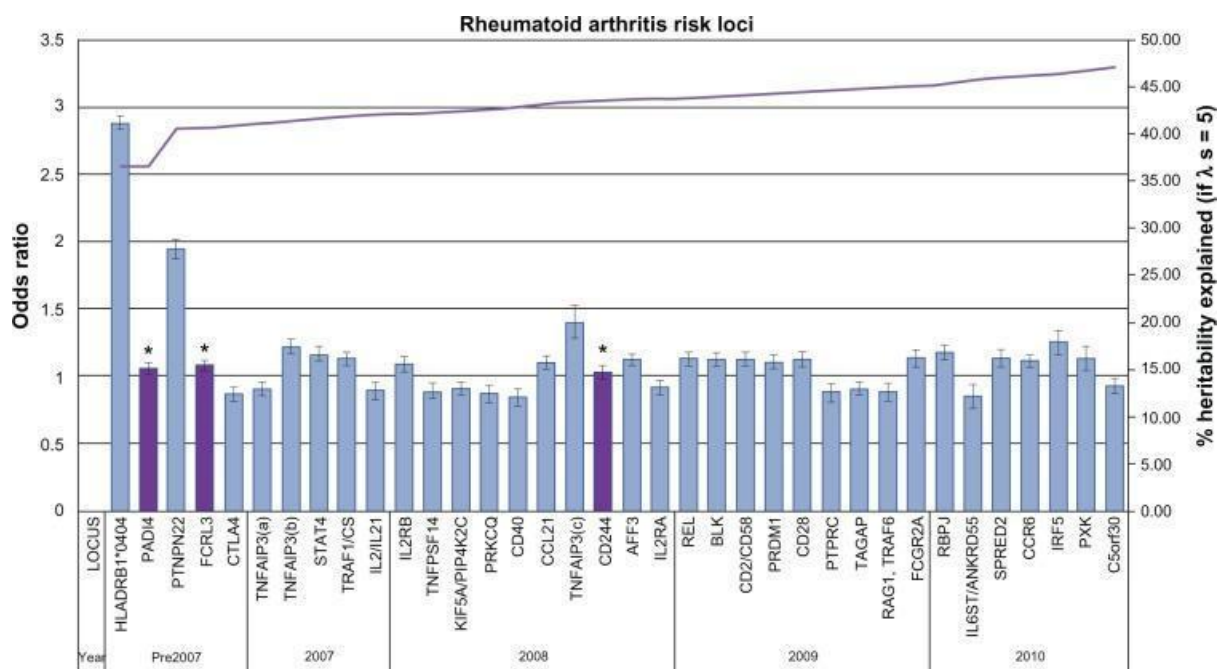


Figure 1: Current rheumatoid arthritis (RA) genetic risk loci (taken from [McAllister et al., 2011](#))

La PR est généralement associée à des preuves sérologiques d'auto-immunité systémique, comme l'indique la présence d'auto-anticorps dans le sérum et le liquide synovial. Le premier auto-anticorps dans la PR, le RF, a été décrit par Waaler en 1940 et est dirigé contre la région Fc des IgG. Les autoantigènes qui comprennent un large éventail de composants du cartilage, de protéines de stress, d'enzymes, de protéines nucléaires et de protéines citrullinées, ciblés par un certain nombre d'autoanticorps, sont finalement trouvés dans la PR. Cela démontre que la PR est caractérisée par des auto-réactivités accumulées dans les cellules B et T.

La pathogénèse de la progression de la PR est médiée par des médiateurs qui activent différentes voies cellulaires. Ces médiateurs interagissent avec les récepteurs membranaires et déclenchent des cascades de signalisation. Après l'interaction ligand-récepteur, le stimulus en amont se propage par une série de réactions couplées de la membrane plasmique au cytoplasme, pour réguler les 10

facteurs clés qui sont responsables des différents phénotypes cellulaires. La PR comprend les principaux stimuli en amont suivants : cytokines et chimiokines, facteurs de croissance, TLR conduisant à l'activation des voies en aval qui comprennent : la voie JAK-STAT, la voie MAPK, la voie PI3K-AKT.

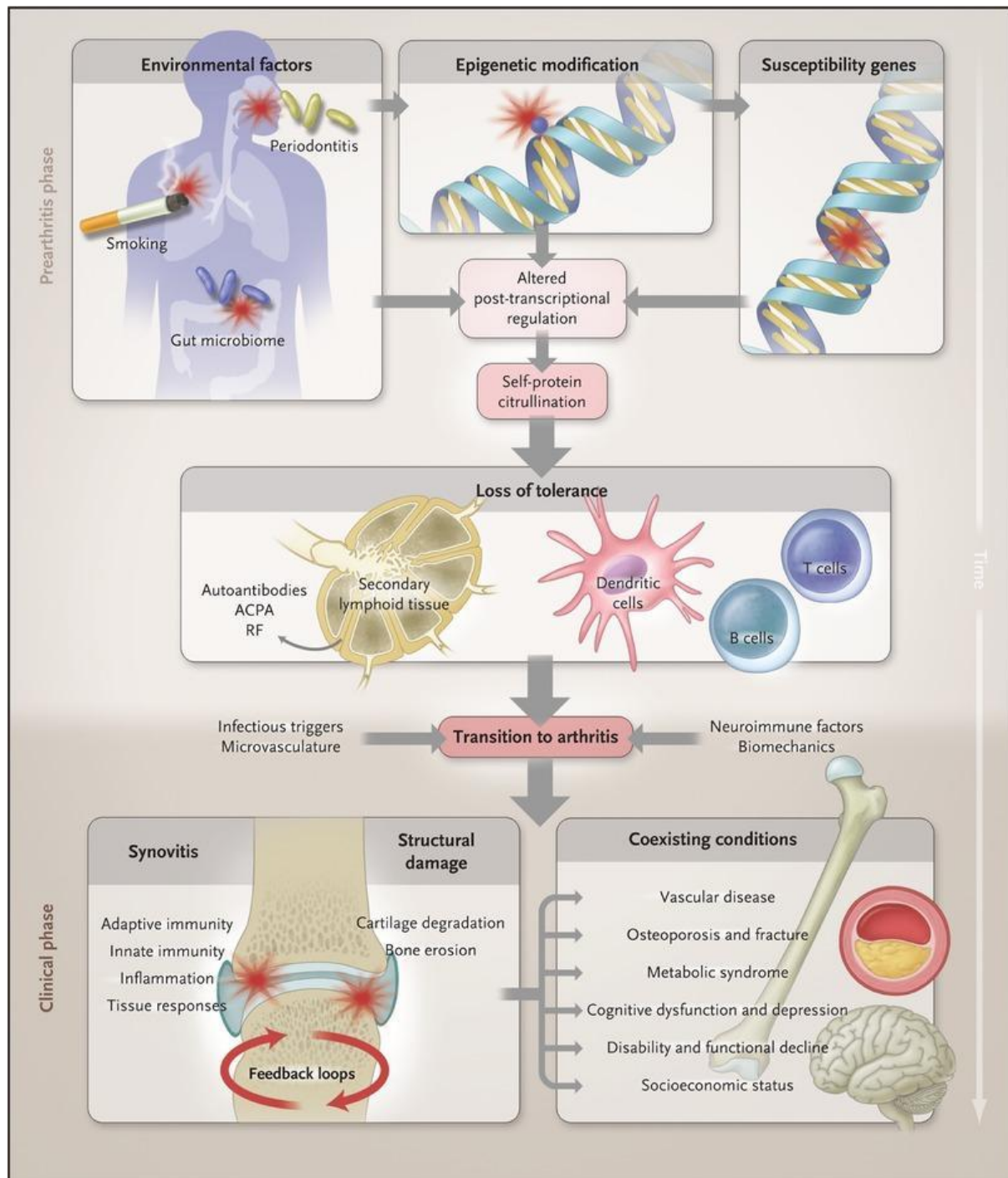


Figure 3: Multistep progression towards RA disease development (figure taken from (McInnes & Schett, 2011))

Les biomolécules des systèmes vivants ne fonctionnent pas individuellement mais interagissent entre elles en formant des réseaux biomoléculaires. Par exemple, une maladie est rarement la conséquence d'une anomalie dans un seul gène mais reflète les perturbations ou les dysfonctionnements des réseaux biologiques complexes qui relient les tissus et les systèmes d'organes. Afin d'organiser et d'analyser cette complexité, il est urgent de développer de nouvelles méthodologies permettant d'obtenir des informations grâce à une vision plus globale.

La biologie systémique est l'étude au niveau des systèmes, qui permet de déchiffrer la complexité de la maladie, son principe fondamental d'initiation et de progression. Elle est considérée comme une approche analytique puissante, qui considère un organisme vivant comme un réseau interactif et dynamique de protéines, de gènes et de réactions biochimiques. L'une des approches initiales communes à cet égard est le domaine de la biologie en réseau qui comprend la construction de réseaux biologiques en exploitant les données expérimentales en constante expansion. La disponibilité de grands ensembles de données en réseau et la capacité de calcul abordable ont conduit au développement d'algorithmes bioinformatiques et d'approches de biologie computationnelle pour analyser les données afin d'offrir des aperçus des dynamiques biologiques. L'étude des systèmes biologiques du point de vue des réseaux a récemment attiré beaucoup d'attention sous la forme de cartes d'interactions présentant les liens entre les protéines (gènes ou neurones), les phénotypes qui leur sont associés (maladies) et les facteurs environnementaux correspondants (médicaments). La modélisation informatique peut être utilisée pour fournir un réseau exécutable et dynamique qui peut révéler des propriétés cachées et prendre en compte les comportements émergents au niveau des systèmes grâce à des simulations et des perturbations *in silico*. Un cycle de construction de modèles, de simulations et de validation expérimentale des prédictions peut contribuer aux approches diagnostiques et thérapeutiques actuelles de la médecine.

La construction de modèles mathématiques implique de construire la structure du modèle, de choisir des expressions mathématiques pour caractériser les relations entre ses composantes, de trouver les valeurs des paramètres et les conditions initiales, et de réaliser des simulations numériques et d'autres analyses mathématiques qui peuvent à la fois reproduire les observations et conduire à des prédictions. De nombreuses méthodes de modélisation basées sur la représentation temporelle et la valeur variable sont en cours d'élaboration pour modéliser et simuler les réseaux moléculaires et génétiques et pourraient être classées de manière générale en deux catégories : les approches quantitatives et qualitatives. Les modèles quantitatifs sont basés sur l'application de la théorie des systèmes à la cinétique chimique et ont été utilisés pour décrire les réseaux métaboliques, les voies de signalisation et la régulation des gènes. La modélisation logique qualitative est basée sur l'idée qu'une variable peut prendre un nombre discret d'états ou de valeurs (deux dans le cas des modèles booléens) et que l'état d'une variable est décidé par une combinaison logique des états d'autres variables.

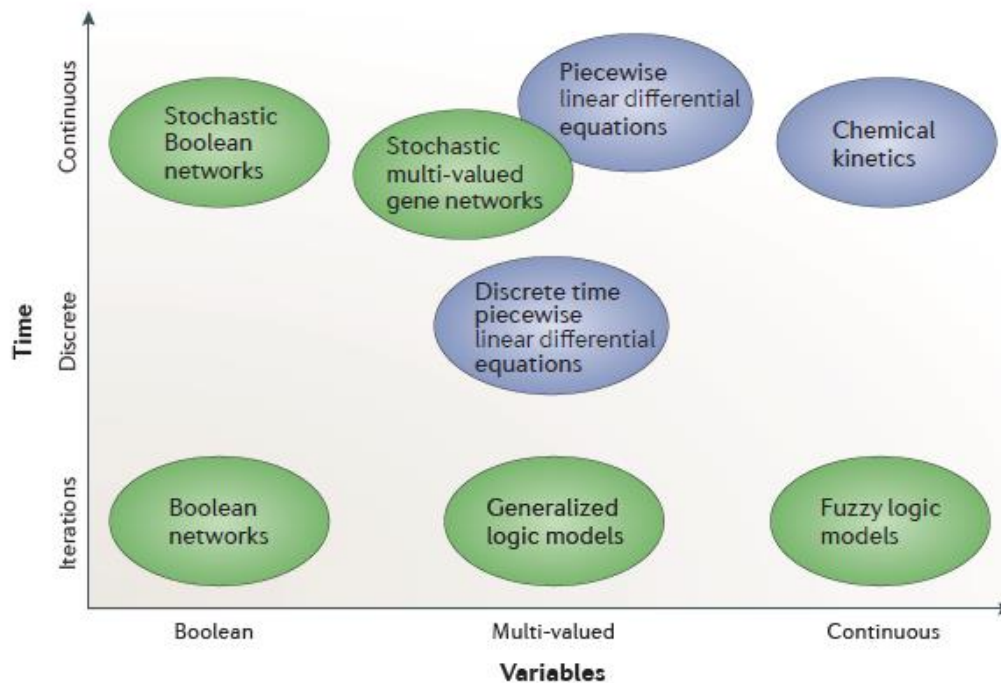


Figure 1.7: Various modelling approaches based on time representation and variable values (taken from (Le Novère, 2015)).

Bien que la modélisation cinétique quantitative fournisse des prédictions plus précises, le manque de paramètres cinétiques entrave son utilisation dans de nombreuses situations, ce qui est abordé par d'autres approches comme la modélisation logique. Dans un modèle logique, chaque composant est associé à une variable discrète, qui est une abstraction logique (souvent booléenne, c'est-à-dire binaire) de son niveau d'activité (ou de concentration). Ce cadre repose sur son évolutivité (des modèles logiques de quelques centaines de composants ont été simulés) ainsi que sur sa nature qualitative, car les paramètres cinétiques et autres connaissances précises sur les mécanismes moléculaires ne sont pas nécessaires. Un modèle booléen fournit une représentation qualitative d'un système. Chaque variable booléenne ne peut prendre que deux valeurs possibles désignées par 1 (ON) ou 0 (OFF) correspondant aux valeurs logiques FAUX et VRAI. ON et OFF représentent l'état d'une entité biologique correspondant à la variable binaire. Cela indique si un gène est exprimé ou non, si un facteur de transcription est actif ou inactif, ou si la concentration d'une molécule est supérieure ou inférieure à un certain seuil. Dans les modèles booléens, l'état futur d'un nœud est déterminé sur la base d'une déclaration logique sur l'état actuel de ses régulateurs. Cette déclaration, appelée règle booléenne (fonction), est généralement exprimée par les opérateurs logiques ET, OU et NON. Les modèles booléens de réseau peuvent être projetés comme un graphe dirigé où les nœuds A, B et C correspondent aux variables booléennes reliées par des arêtes.

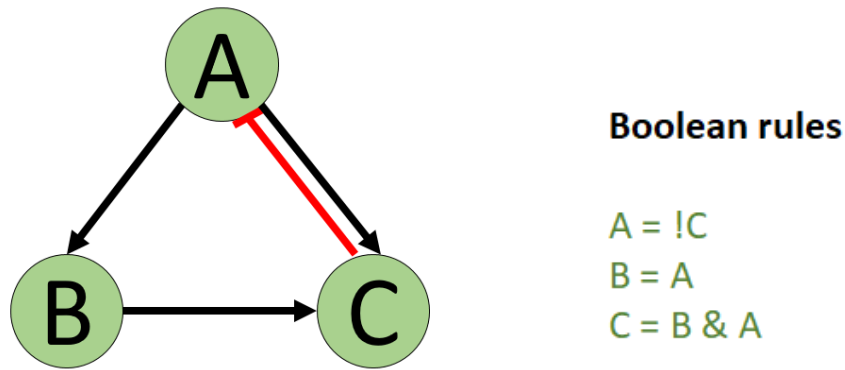


Figure 8: A simple Boolean network model: A simple regulatory network containing three nodes A, B and C is shown. The edges with sharp arrows represent activation (positive effects) while the edges with blunt arrows indicate inhibition (negative effects).

Pour comprendre les détails mécaniques des processus biologiques sous-jacents de la maladie complexe qu'est la PR, nous devons systématiquement rassembler et analyser les informations disponibles dans la littérature, les ensembles de données, et les bases de données. L'objectif de ce travail de doctorat est de faire la lumière sur les mécanismes moléculaires impliqués dans la pathogenèse de la PR en utilisant des approches, des outils et des techniques de pointe en biologie systémique. Le premier objectif du travail est de résumer les connaissances biologiques actuelles (expression des gènes, voies de signalisation, phénotypes cellulaires) concernant la PR en une carte détaillée d'interactions moléculaires. La construction d'une carte de la PR à la pointe de la technologie nécessite une analyse exhaustive de la littérature, une réévaluation des tentatives publiées précédemment, ainsi que des critères de conservation stricts et, surtout, des conseils d'experts afin de limiter les faux positifs en se concentrant sur la maladie et la spécificité humaine. Les cartes d'interactions moléculaires peuvent servir de base de connaissances autonomes, ou elles peuvent être utilisées comme un échafaudage pour la construction de modèles informatiques. Un aspect important est également l'utilisation de normes communautaires telles que le SBGN, qui est une représentation graphique standard destinée à favoriser le stockage, l'échange et la réutilisation efficaces des informations sur les voies de signalisation, les réseaux métaboliques et les réseaux de régulation des gènes parmi les communautés de biochimistes, de biologistes et de théoriciens. Ces cartes résument alors les connaissances actuelles sur les voies biologiques dans une représentation de description de processus, tout en tenant compte du plus grand nombre possible de détails mécanistiques. C'est la première tentative de construction d'une carte des interactions moléculaires pour la PR en utilisant presque exclusivement des données spécifiques à l'homme et à la maladie, et en utilisant la norme SBGN. Le second objectif du projet est de développer une approche de modélisation informatique pour fournir un réseau dynamique exécutable qui peut révéler des propriétés cachées et prendre en compte les comportements émergents au niveau du système par le biais de simulations et de perturbations *in silico*. Parmi les différents types de cellules impliquées dans la PR, les FS jouent un rôle crucial dans la persistance des caractéristiques destructrices de la maladie. Il est démontré qu'ils expriment des cytokines immunomodulatrices,

diverses molécules d'adhésion et des enzymes de modélisation de la matrice. En outre, les fibroblastes de la PR présentent des taux de prolifération élevés et un phénotype résistant à l'apoptose. Ces cellules peuvent également se comporter comme des moteurs primaires de l'inflammation, et les thérapies dirigées contre les FS dans la PR pourraient devenir une approche complémentaire aux thérapies dirigées par le système immunitaire.

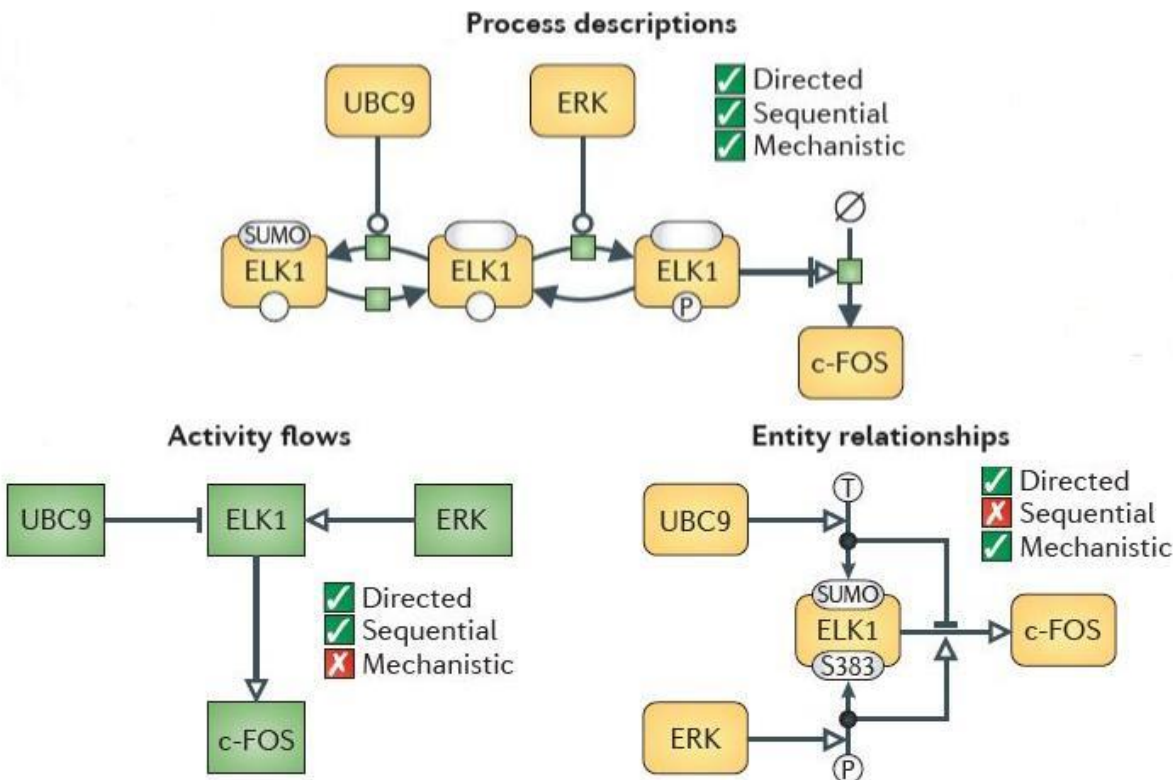


Figure 5 : Three different types of SBGN networks used to represent biological processes (taken from (Le Novère, 2015))

Notre objectif est de construire un modèle à grande échelle pour étudier le comportement des FS de la PR afin de mieux comprendre les mécanismes qui contrôlent leur phénotype agressif. Il s'agit, à notre connaissance, de la première tentative de construction d'un modèle dynamique à grande échelle pour ce type d'étude. Le troisième objectif du projet est un effort pour automatiser le passage d'une représentation statique des mécanismes de la maladie à un modèle dynamique développant un cadre pour la conversion des graphes et l'inférence de formules logiques basées sur la topologie et la sémantique encodées sur la carte moléculaire. La construction d'un modèle dynamique exécutable, soit à partir de cartes/modèles moléculaires préexistants, soit de manière autonome, est un processus qui prend du temps. Le cadre proposé pourrait considérablement faciliter le processus de génération de modèles exécutables prêts à l'emploi pour les simulations *in silico* (perturbations/déclenchements). Notre modèle est le premier modèle booléen à grande échelle à être construit de manière entièrement automatisée à partir d'une carte d'interaction moléculaire. En outre, cela ouvre également la voie à la mise à l'échelle de la génération automatique de modèles à grande échelle. Jusqu'à présent, le nombre de nœuds pouvant être inclus dans un modèle biologique dynamique était limité. Cela était conforme aux outils de modélisation

disponibles dont le potentiel de simulation était également limité. Dans cette thèse, nous voudrions remettre en question ce potentiel et augmenter l'échelle des modèles couramment utilisés pour étudier les réseaux biologiques, afin d'élucider les mécanismes de régulation complexes et de donner des réponses à des questions plus sophistiquées qui ne pourraient pas être traitées avec des modèles comprenant un nombre limité de facteurs.

Dans ce sens, une carte complète des interactions moléculaires pour la PR, construite avec le logiciel CellDesigner, a été publiée en 2010. Elle était basée sur des données expérimentales à haut débit combinées avec des informations provenant de la base de données KEGG (<http://www.genome.jp/kegg/pathway.html>). Vingt-huit études publiées ont été utilisées pour la construction de la carte qui comprenait des expériences réalisées dans différents types de cellules et de tissus tels que la cellule mononucléaire du sang périphérique (PBMC), les fibroblastes synoviaux, les tissus synoviaux et les macrophages entre autres. Nous avons utilisé cette carte comme base et l'avons étendue pour créer une carte de la PR basée sur les connaissances les plus récentes. Nous avons transformé la carte de la PR en une base de connaissances interactive de pointe sur la maladie, qui s'interface avec diverses bases de données pour l'annotation du contenu et l'analyse d'enrichissement des résultats expérimentaux. Nous avons également utilisé des outils bioinformatiques tels que BioInfoMiner (<https://bioinfominer.com>) et Cytoscape (<https://cytoscape.org/>) pour l'analyse de la carte de la PR comme un réseau biologique complexe, révélant les aspects topologiques et fonctionnels de la carte.

La carte de la PR illustre graphiquement les voies de signalisation, la régulation de l'expression des gènes, les mécanismes moléculaires et les phénotypes cellulaires impliqués dans la pathogenèse de la maladie. Cette carte implique une recherche documentaire exhaustive, l'extraction d'informations dans les bases de données pertinentes ainsi qu'une mise à jour continue et des conseils d'experts du domaine. Il est important de noter que les interactions illustrées dans le diagramme représentent un modèle graphique codé selon un format standardisé, ce qui rend la carte facile à calculer. Tous les composants de la carte ont au moins deux références PubMed créées manuellement, ce qui donne un total de 353 publications couvrant une période allant de 1975 à 2019.

Nous présentons donc ici une carte des interactions moléculaires à grande échelle pour la PR, qui est à notre connaissance la première carte des maladies normalisée conforme au SBGN (description des processus). Tous les composants et réactions sont annotés en utilisant uniquement la PR et les études spécifiques à l'homme. Afin de limiter les faux positifs, les conseils d'experts sont intégrés et les normes SBGN ont été utilisées pour la représentation de la carte afin d'assurer sa réutilisation. La carte de la PR que nous présentons ici comprend des informations provenant de 353 publications évaluées par des pairs et présentant 506 espèces, 446 réactions et 8 phénotypes. Les espèces de la carte sont classées en 303 protéines, 61 complexes, 106 gènes, 106 entités ARN, 2 ions et 7 molécules simples.

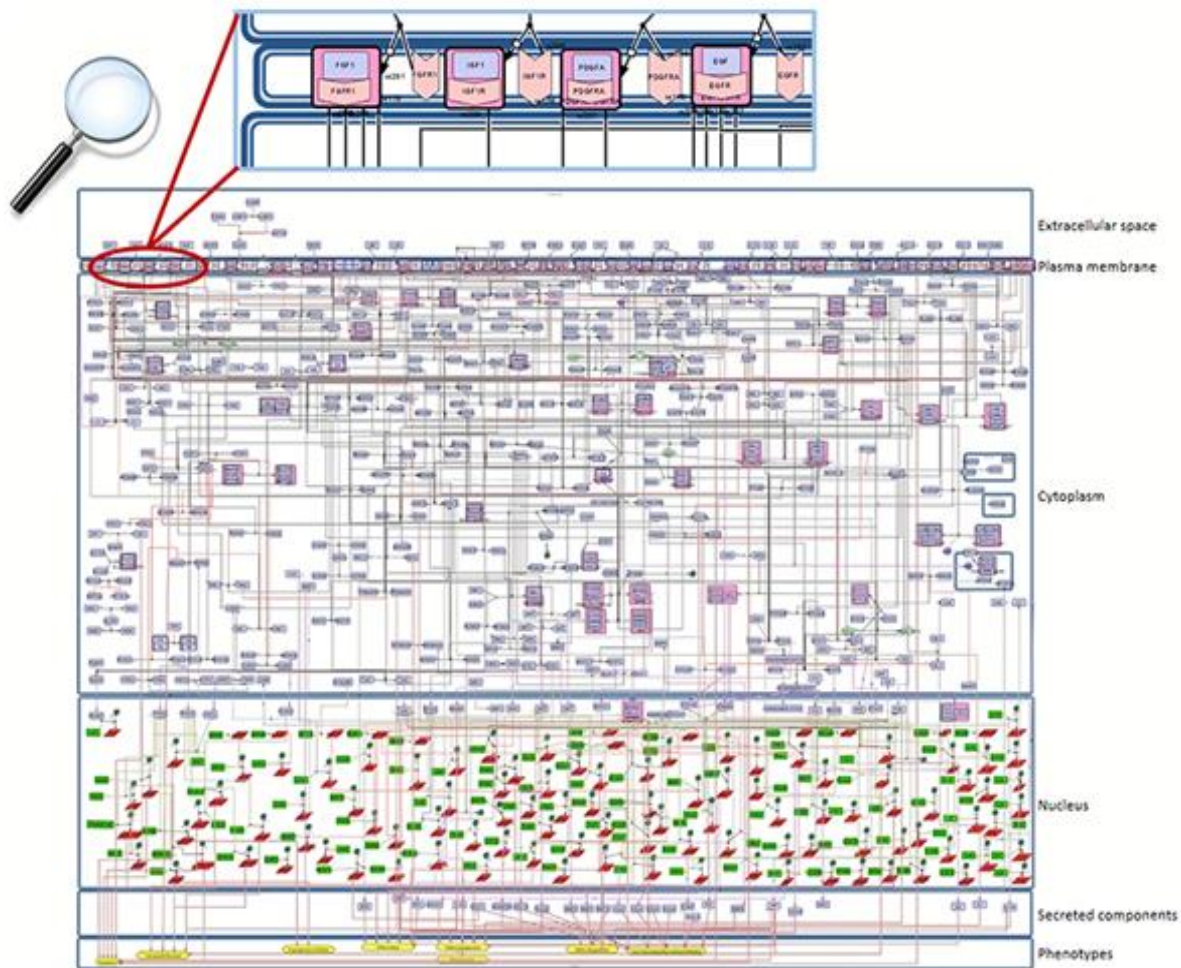


Figure. La carte de la PR

La carte de la PR peut également être utilisée comme une base de connaissances interactive, en utilisant la plateforme MINERVA, et servir de modèle pour la superposition de plusieurs ensembles de données. La visualisation des données expérimentales peut en effet aider à mettre en évidence certains aspects du processus biologique affecté et à rendre plus évidentes les différences entre les conditions expérimentales. La visualisation des résultats de l'analyse de l'expression différentielle de trois ensembles de données sur l'expression génique des tissus synoviaux de la PR a montré un enrichissement dans tous les phénotypes cellulaires mais pas dans l'apoptose. Cette constatation est conforme au fait que les fibroblastes, qui constituent un pourcentage important des synoviocytes de la PR, présentent un phénotype résistant à l'apoptose.

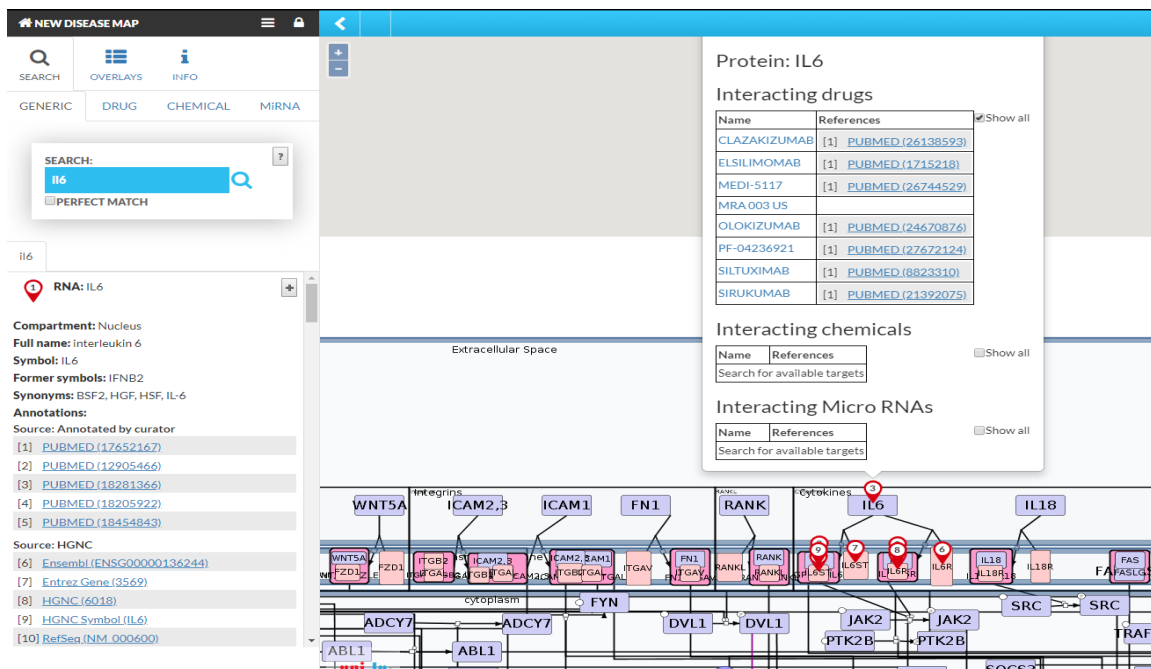
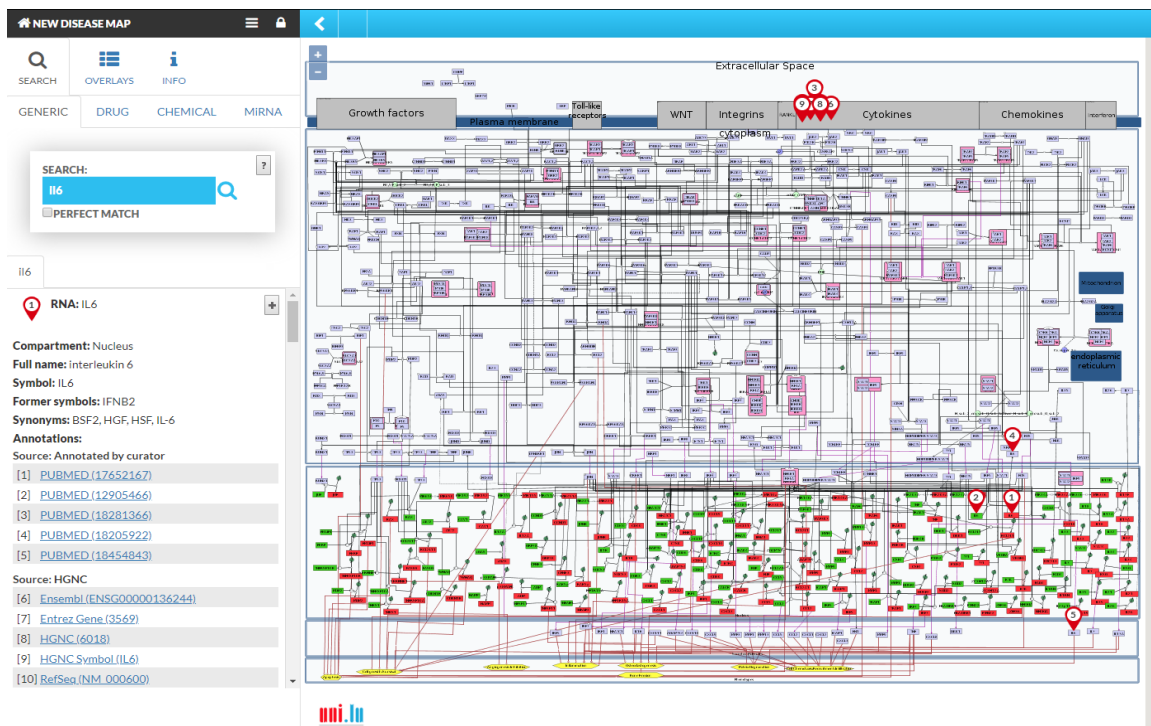




Figure : Minerva

Nous avons effectué une analyse fonctionnelle et une priorisation des gènes à l'aide du logiciel BioInfoMiner. Les gènes qui sont les mieux classés dans cette analyse sont associés à de nombreux processus systémiques et sont considérés comme des nœuds dans le réseau sémantique. En plus de la priorisation, une analyse pharmacogénomique est fournie puisque les hubs proposés sont considérés comme des cibles médicamenteuses putatives. Les résultats de l'analyse utilisant les termes GO et PHO ont révélé des acteurs connus de la PR, dont la plupart ont déjà été utilisés comme cibles de médicaments, ce qui démontre que la carte de la PR comprend des facteurs bien caractérisés et saisit la plupart des processus systémiques pertinents impliqués dans la maladie. L'analyse topologique peut révéler des caractéristiques structurales sous-jacentes de la carte de la PR comme des parties non connectées du réseau ou des nœuds importants (nœuds bien connectés) qui sont autrement difficiles à percevoir dans les réseaux à grande échelle. L'analyse topologique réalisée dans le cadre de cette étude a révélé des parties connectées et non connectées du réseau. Ce résultat reflète notre connaissance fragmentée d'une part, mais aussi l'utilisation de critères stricts pour les nœuds inclus dans la carte : interactions validées expérimentalement dans au moins deux études publiées, utilisation de données d'origine strictement humaine et spécifiques à la maladie. La carte de la PR, fruit de collaborations interdisciplinaires entre cliniciens, biologistes et bioinformaticiens, est accessible sur ramap.elixir-luxembourg.org.

La construction de modèles dynamiques à grande échelle peut être un travail long et fastidieux qui nécessite non seulement la construction du graphique réglementaire mais aussi l'écriture et le réglage des formules logiques. Nous avons développé un outil, CaSQ, visant à faciliter la construction de modèles booléens à grande échelle, en tirant parti des similitudes partagées entre les cartes d'interaction moléculaire et les modèles dynamiques. Dans le cadre proposé, nous utilisons des normes de biologie systémique pour la construction de modèles (SBML-qual), de

sorte que l'outil CaSQ puisse être interopérable avec d'autres outils et logiciels de modélisation. Pour l'inférence des formules logiques, nous avons basé nos hypothèses sur la topologie et la sémantique des cartes moléculaires. Plus précisément, nous avons décidé d'aborder le processus de conversion en utilisant principalement des fonctions logiques OR sur AND, de sorte qu'une cible est activée si l'une des réactions qui la produit est activée, une réaction est activée si tous les réactifs sont activés, tous les inhibiteurs sont désactivés et l'un des catalyseurs est activé. L'idée qui sous-tend cette hypothèse est que, très rarement, nous disposons d'informations exactes sur la nécessité de la présence de deux ou plusieurs activateurs pour une cible. Même si une synergie est définie, très souvent une activation relative peut se produire même par la présence d'un seul activateur. En outre, le nombre d'événements pour lesquels nous disposons de telles informations est nettement inférieur à celui des événements incertains et le réglage des règles à la main devrait être un processus rapide. Grâce à l'utilisation de CaSQ, comme le montre cette étude, nous pouvons maintenant obtenir des modèles booléens à grande échelle qui peuvent être exécutés à l'aide de logiciels de modélisation populaires capables d'importer des fichiers de qualité SBML. Dans ce travail, pour comparer les performances et la précision de l'outil, nous avons comparé les nœuds communs entre le CaSQ déduit et les modèles construits manuellement, leur capacité à reproduire des scénarios biologiques en effectuant des simulations, et enfin, nous avons effectué une comparaison des états stables, lorsque cela était possible. Nous avons également effectué des simulations pour voir si les modèles interférés par le CaSQ pouvaient reproduire une partie de la dynamique du système original. L'étape suivante a consisté à effectuer une analyse logique de l'état stable.

Pour cela, nous avons utilisé GINsim, un logiciel puissant de modélisation logique. Le but était de voir si, dans les états stables du modèle interférentiel CaSQ, nous pouvions récupérer les états stables du modèle construit manuellement et publié. Il faut noter que CaSQ déduit des règles booléennes préliminaires, de sorte que le modélisateur doit encore affiner le modèle et trouver les meilleures règles logiques pour reproduire les données avec précision. Bekkar et al, 2018 montrent que les modèles logiques avec une curation humaine ajoutée ont de meilleures performances que les modèles où les règles sont extraites automatiquement d'une topologie donnée. Comme le montrent les résultats, l'outil CaSQ produit des modèles qui sont largement en accord avec le modèle qu'un modélisateur humain construirait, ce qui accélère de façon impressionnante le temps de construction du modèle. Ce travail a également motivé le travail communautaire, car il a abordé les questions de réutilisation des modèles, d'utilisation des formats standard de la biologie systémique et d'interopérabilité entre différents outils ayant des fonctionnalités complémentaires. Comme démontré, notre méthode est évolutive, et les modèles SBML-qual à grande échelle produits par CaSQ peuvent être importés dans Cell Collective et conserver la mise en page et les annotations. L'objectif est de proposer un pipeline transparent pour la production de modèles booléens exécutables à partir de cartes d'interactions moléculaires qui peuvent être analysées en profondeur à l'aide de divers outils de modélisation informatique. L'outil CaSQ peut jouer le rôle d'un pont réunissant deux communautés distinctes, les conservateurs et les modélisateurs, pour produire des modèles annotés interopérables de meilleure qualité, plus précis et réutilisables.

Pour l'objectif final du doctorat, nous avons mis à notre connaissance nos efforts pour créer un modèle booléen à grande échelle pour les FS de PR, le premier modèle exécutable pour ce système. Nous utilisons la carte de la PR, une source de connaissances de haute qualité basée sur la curation humaine, et les réalisations techniques décrites dans les chapitres précédents pour construire un modèle axé sur des résultats cellulaires spécifiques, à savoir l'apoptose, la prolifération cellulaire, la dégradation de la matrice, l'érosion osseuse et l'inflammation. Nous avons dû faire un

compromis dans le choix des phénotypes, car la taille du modèle produit était déjà importante. Nous avons fait ce choix en fonction de nos principales questions biologiques, et plus particulièrement de notre intérêt pour l'étude de la résistance à l'apoptose, de la contribution à l'inflammation et des résultats liés aux dommages structuraux. Cependant, le phénotype que nous aimerions inclure à l'avenir est la chimiotaxie et le recrutement cellulaire, car les FS de PR sont connus pour sécréter un certain nombre de cytokines et de chimiokines qui fonctionnent comme des appels de signalisation vers d'autres cellules qui s'infiltrant dans les articulations. Pour faire face à la complexité et aux défis des simulations à grande échelle, nous utilisons une approche modulaire pour construire notre modèle et les sous-modules qui le créent. Cette approche nous a permis d'étudier des parties plus petites des systèmes, et en même temps d'évaluer le comportement au niveau des modules et du modèle entier.

Les simulations en temps réel des modules et de l'ensemble du modèle ont révélé des incohérences et d'éventuelles interactions manquantes, notamment pour les modules d'apoptose et de dégradation de la matrice. Une observation des simulations en temps réel est que les comportements des modules individuels et du modèle entier pour les scénarios testés étaient cohérents. Cela signifie que le comportement observé dans le module a été maintenu intact dans l'ensemble du modèle. Cela pourrait être dû au fait que la majorité des simulations impliquent l'activation ou l'inactivation d'une ou de quelques entrées seulement, mais ce n'était pas un résultat de facto. Nous avons pu calculer des attracteurs pour notre système en utilisant le logiciel BoolNet de R et des heuristiques qui ont contribué à limiter la recherche de l'espace d'état. Nous avons choisi comme conditions initiales soit tous les nœuds mis à zéro, soit tous mis à un. Nous avons calculé les attracteurs pour les deux conditions, pour chaque module individuel et pour l'ensemble du modèle. Leur analyse, préliminaire à l'heure actuelle car ils sont de taille considérable, a confirmé les comportements observés lors des simulations en temps réel, et a également révélé quelques incohérences concernant le comportement de certains nœuds (par exemple AKT2 dans l'inflammation). Une observation intéressante concerne les oscillations entre les entités TP53 et MDM2 dans notre modèle. L'expression abondante de MDM2 dans les FS de PR a été démontrée. MDM2 est le principal régulateur négatif de p53 et, dans les tumeurs, elle contribue à augmenter la prolifération cellulaire. Dans la PR, elle pourrait être un facteur contribuant au phénotype hypoapoptotique des tissus de la muqueuse grâce à sa capacité à réguler à la baisse les niveaux et les effets de p53. Les anomalies de p53 dans la PR pourraient soutenir et accélérer l'inflammation synoviale principalement par l'intermédiaire de l'IL-6, comme l'ont montré des études utilisant des rats Lewis avec une arthrite induite par un adjuvant (AIA). Une analyse approfondie des attracteurs est nécessaire et est actuellement en cours, afin de fournir une vision plus détaillée du comportement du système. L'étape suivante consisterait à examiner attentivement l'état de tous les composants dans des conditions données afin d'évaluer dans quelle mesure le comportement du modèle est cohérent avec la biologie du système. Différentes conditions initiales pourraient également être utilisées pour tester des hypothèses spécifiques. Les probabilités phénotypiques utilisant MaBoSS ont fourni une troisième couche d'analyse des comportements observés, notamment en ce qui concerne la dépendance de l'apoptose et le rôle de la voie WNT dans l'érosion osseuse. D'autres scénarios de simulation devraient être testés afin d'obtenir des informations sur des mécanismes spécifiques.

Au cours de ma thèse, la taille des modèles obtenus, même pour les modules, a remis en question le potentiel d'analyse et de simulation de la plupart des logiciels utilisés pour les modèles logiques. L'un des problèmes que nous avons rencontrés au début était le fait que la plupart des logiciels nécessitent l'utilisation du même logiciel pour la construction et l'analyse des modèles. C'était

également le cas pour Cell Collective. Notre collaboration avec le Dr Tomas Helikar de l'université du Nebraska-Lincoln, aux États-Unis, a conduit à la mise en œuvre de l'importation de la qualité SBML pour Cell Collective, ce qui a grandement facilité l'analyse. Le Dr Helikar nous a également aidé à récupérer les références de nos modèles encodés dans les identificateurs MIRIAM, dans la plate-forme Cell Collective, ce qui nous a permis de créer des modèles annotés instantanément qui peuvent être publiés sans avoir à refaire le travail de conservation. Tout au long de cette thèse, les développeurs de Cell Collective ont ajusté la plateforme pour satisfaire nos demandes toujours croissantes et ont soutenu sans réserve notre quête de modélisation à grande échelle. Nous avons également travaillé en étroite collaboration avec le Pr. Denis Thieffry (ENS Paris) et le Dr. Aurelien Naldi (ENS Paris), pour surmonter les problèmes d'interopérabilité concernant l'importation de nos fichiers SBML-qual dans GINsim. Lorsque la taille et la complexité des modèles ont remis en question la capacité de GINsim, bioLQM a été utilisé pour effectuer des réductions. Le Dr. Laurence Calzone (Institut Curie Paris), a également fourni des conseils utiles pour l'utilisation de MaBoSS et les réglages nécessaires au fonctionnement du logiciel. Enfin, je voudrais mentionner que nous avons eu le soutien du Dr. Sylvain Soliman (INRIA), qui, entre autres, a aidé à nommer les différents composants du modèle, facilitant ainsi l'identification des gènes, des ARN et des protéines ayant un nom commun. Sans les efforts collectifs et le soutien de la communauté, la plupart des travaux présentés n'auraient pas été possibles. Maintenant que l'interopérabilité est réalisée et que le cadre carte-modèle est établi, on peut espérer que l'analyse en aval des réseaux booléens biologiques à grande échelle sera plus facile et moins longue d'un point de vue technique, ce qui laissera plus de temps pour se concentrer sur les questions biologiques en jeu.

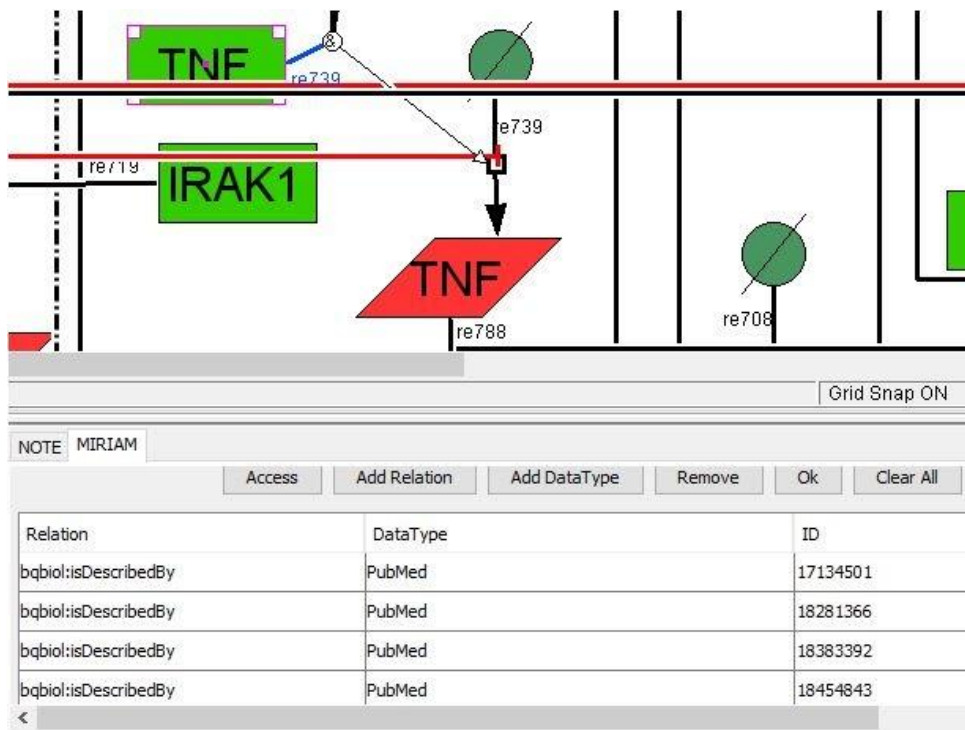


Figure 6: CellDesigner Annotation with MIRIAM

Pour conclure, les systèmes vivants ne peuvent être compris en étudiant uniquement leurs parties individuelles. Avec une production sans cesse croissante d'ensembles de données biologiques, accélérée par le séquençage des génomes et les techniques omiques à haut débit, l'objectif général de la recherche biomédicale sur les maladies complexes devrait être progressivement déplacé d'une analyse essentiellement en régime permanent au niveau moléculaire à un niveau de biologie systémique capturant le comportement dynamique caractéristique du système. La biologie systémique est une approche holistique qui permet d'étudier la vue d'ensemble des systèmes biologiques et de leur organisation. Une tentative de définition des qualités de la biologie des systèmes a été proposée, avec trois caractéristiques principales. Premièrement, la diversité qui se réfère à la compréhension biologique que chaque interaction d'un composant apporte au système. Deuxièmement, la simplicité qui fait référence à l'approche réductionniste en décomposant le système en descriptions simples. Troisièmement, la complexité, qui renvoie à la compréhension de l'interaction complexe d'un réseau moléculaire. Ces trois qualités font de la biologie systémique un domaine interdisciplinaire par nature, qui combine l'informatique, l'informatique et les mathématiques avec la biologie.

Entre autres applications, il s'est révélé être un outil analytique puissant pour comprendre les interactions dynamiques qui sont au cœur des maladies complexes. Pour résumer, dans ma thèse :

a) J'ai utilisé mes connaissances préalables pour construire un réseau spécifique aux FS de la PR. Pour récapituler tout ce qui est connu et publié (dans diverses sources) sur la maladie, une carte mondiale, entièrement complète et annotée, spécifique à la PR a été construite sur la base d'une recherche documentaire exhaustive, d'un traitement humain et d'une validation par des experts du domaine. Cette carte présente les interactions impliquées dans la PR provenant de divers types de cellules. Grâce aux annotations étendues de chaque entité et de chaque réaction incluses dans la carte, et aux fonctionnalités avancées offertes par la plateforme MINERVA, l'utilisateur peut opter pour des interactions spécifiques aux cellules et extraire les réseaux correspondants spécifiques aux cellules. Si de telles cartes de connaissances sont très instructives, elles pourraient être pleinement exploitées en les traduisant en objets dynamiques et en modèles mathématiques prédictifs.

b) J'ai travaillé pour ajouter une couche dynamique au réseau de manière automatisée. Jusqu'à présent, la construction des modèles dynamiques était manuelle, sauf pour quelques-uns. Pendant mes études doctorales, un effort pour ajouter une couche dynamique sur cette carte descriptive a été exploré, et l'approche choisie pour étudier la carte la PR de manière dynamique est le formalisme logique (booléen), pour sa simplicité et son absence de paramètres cinétiques. En collaboration avec le Dr Soliman, (Lifeware INRIA, Saclay) nous avons développé CaSQ (CellDesigner as SBML-Qual), <https://lifeware.inria.fr/~soliman/post/casq/>), un outil pour l'inférence automatisée de modèles booléens à grande échelle, sans paramètres, à partir de cartes d'interactions moléculaires basées sur la topologie et la sémantique des réseaux. Il s'agit, à notre connaissance, du premier outil qui produit des modèles logiques exécutables de centaines de nœuds (jusqu'à plusieurs centaines), dans un format standard SBML pour la description des modèles, qui peuvent être simulés et analysés plus avant à l'aide d'outils de modélisation populaires. Dans ce cadre, des formules logiques préliminaires pour le modèle sont déduites automatiquement selon des règles et des contraintes prédéfinies. Ces règles se sont avérées suffisamment générales pour couvrir divers scénarios de représentation des connaissances biologiques, mais n'ont pas été testées systématiquement pour évaluer la robustesse du modèle.

c) Le problème de l'analyse des modèles booléens à grande échelle constitue un défi sur le terrain. La taille accrue du modèle en termes de nombre de nœuds et la complexité due aux diaphonies de signalisation et à la présence de nombreuses boucles de rétroaction ne permettent pas une analyse dynamique simple. Pour faire face à la complexité, j'ai utilisé une approche modulaire qui m'a permis d'étudier chaque phénotype fonctionnel des FS de PR séparément et de le comparer avec le système fusionné. Pour ce faire, j'ai utilisé la carte globale et le plugin de flux de la plateforme MINERVA pour me concentrer sur les sous-réseaux pertinents. L'idée de modules qui partent d'un phénotype et incluent tous les régulateurs possibles en amont a simplifié la charge de simulation des relations entrée-sortie et a aidé à mieux comprendre le système des FS de PR.

d) Outre la modularité, qui m'a aidé à faire face à la complexité du système en utilisant la stratégie du "diviser pour mieux régner", j'ai également essayé de regrouper les entrées du modèle qui étaient nombreuses et qui contribuaient à l'explosion des calculs lorsque j'essayais de faire une analyse dynamique. Bien que le regroupement des entrées entraîne une certaine perte d'informations (c'est-à-dire que je ne pouvais pas effectuer de simulations *in silico* pour le TNF et l'IL6 séparément car ils étaient regroupés sous forme de cytokines), cela a rendu le calcul des probabilités phénotypiques accessible et a contribué à une meilleure compréhension du système.

e) J'ai utilisé des normes de biologie systémique pour la conservation, la notation graphique et la modélisation afin de promouvoir l'utilisation de normes communautaires, la transparence des résultats scientifiques, l'interopérabilité et la réutilisation des fichiers et du contexte scientifique, respectivement. J'ai utilisé le SBGN pour la construction de la carte RA, les identifiants MIRIAM pour la conservation, les identifiants uniques pour les gènes et les protéines, et les fichiers SBML qual pour la modélisation dynamique. Tous mes travaux sont ou seront accessibles en accès libre, ce qui favorise la science ouverte.

f) J'ai également eu la chance de faire du travail communautaire et d'apprendre de cette expérience. Je suis membre du consortium Disease Map et j'ai contribué activement à l'initiative COVID-19 Disease Map pour la carte moléculaire de processus Apoptose (voir chapitre 6 dans la thèse - contributions à la communauté).

Orientations futures:

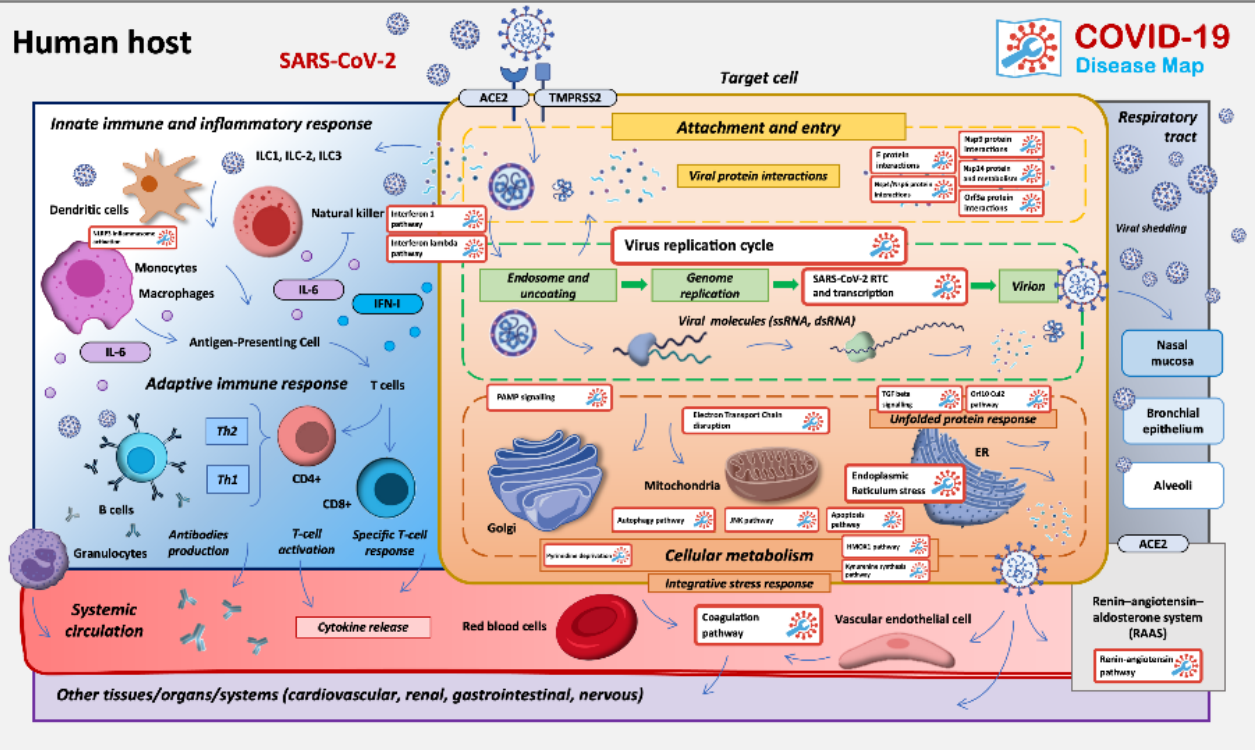
a) Il est difficile d'estimer la robustesse d'un modèle lorsque les hypothèses pour l'inférence des règles logiques sont modifiées (par exemple, utiliser les portes ET au lieu de OU pour les formules logiques). La robustesse du modèle pourrait être estimée comme la capacité du modèle à reproduire un scénario biologique bien établi (validé expérimentalement), basé uniquement sur les règles déduites (sans réglage manuel supplémentaire). Une étude approfondie de l'impact d'une modification des règles de traduction du modèle permettra de mieux comprendre les règles et l'effet de la topologie sur la dynamique du modèle. Cet impact pourrait être résumé en probabilités phénotypiques de résultats fonctionnels selon différents scénarios et ensembles de règles. Les résultats pourraient ensuite être systématiquement confrontés aux données expérimentales à petite échelle et aux connaissances des experts encodées dans la carte de l'AR afin d'identifier l'ensemble de règles de traduction le plus robuste/fiable.

Human host

SARS-CoV-2



COVID-19
Disease Map



- b) La présence de nombreux intrants dans les modèles inférés est le résultat combiné des règles de traduction de carte à modèle et de la structure de la carte qui conduit à la création de pseudo-intrants. Ces pseudo-intrants augmentent le coût de calcul de l'analyse, et une stratégie qui pourrait aider à limiter leur présence/impact (c'est-à-dire en fixant leurs valeurs) dans les modèles obtenus faciliterait l'analyse en aval.
- c) L'ambition de ce travail est de mettre en place un cadre qui puisse faciliter l'identification de nouvelles cibles thérapeutiques pour la PR. L'objectif est de prédire les conditions optimales qui favoriseraient l'apoptose et minimiseraient l'érosion osseuse, la destruction du cartilage et les résultats de l'inflammation. Des tests systématiques de différentes conditions initiales pourraient en outre permettre de prédire les résultats de perturbations spécifiques, telles que des effets uniques ou combinés, simulés avec le modèle en forçant ou en supprimant systématiquement l'activité de chaque gène/protéine. L'objectif final est de mieux comprendre le mécanisme de dégradation du cartilage et de l'os, deux symptômes débilissants majeurs de la PR, et de proposer une stratégie qui pourrait aider à bloquer, voire à inverser ces résultats.
- d) Pour obtenir un modèle affiné et mieux formé, les données issues d'expériences à petite échelle et les ensembles de données Omic publiées sur les fibroblastes de la PR (transcriptomique, RNAseq, RNAseq mono-cellulaire) devraient être utilisés pour augmenter la spécificité des cellules et de la maladie, ainsi que pour définir des signatures phénotypiques (biomarqueurs) qui pourraient être testées par rapport à des états stables.
- e) Les approches de test unitaire et de propagation de la valeur telles que celles décrites dans Hernandez et al, 2020, pourraient également être appliquées pour vérifier le comportement du modèle au niveau local.
- f) Une autre application potentielle consisterait à identifier les points du système FLS de l'AR qui sont susceptibles d'être drogués et qui ne sont pas seulement visés par les médicaments anti-rhumatismaux, et à effectuer des simulations avec des perturbations combinées. Cette approche pourrait permettre d'identifier les traitements thérapeutiques combinés ou de réorientation des médicaments et d'évaluer l'impact sur les phénotypes.

Enfin, les expériences in vitro seraient un complément idéal de ces travaux. Les prédictions du modèle pourraient être utilisées pour approfondir l'étude du mécanisme de la maladie au niveau intracellulaire, en ciblant la voie/les facteurs proposés et en mesurant l'apoptose crosstalks cellulaires, en évaluant l'impact des résultats fonctionnels des FS de la PR sur les cellules du système immunitaire et d'autres cellules voisines telles que les chondrocytes, responsables des lésions du cartilage et de l'érosion osseuse qui s'ensuit. Des expériences in vitro pourraient également être utilisées pour tester les prévisions concernant les traitements combinés.