



HAL
open science

Analyse de réseaux de régulation par approches de coloration de graphes dans le cadre du myélome multiple

Bertrand Miannay

► To cite this version:

Bertrand Miannay. Analyse de réseaux de régulation par approches de coloration de graphes dans le cadre du myélome multiple. Bio-Informatique, Biologie Systémique [q-bio.QM]. Ecole centrale de Nantes, 2017. Français. NNT: . tel-01679863v2

HAL Id: tel-01679863

<https://hal.science/tel-01679863v2>

Submitted on 10 Jan 2018 (v2), last revised 10 May 2021 (v3)

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

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

Thèse de Doctorat

Bertrand MIANNAY

*Mémoire présenté en vue de l'obtention du
grade de Docteur de l'École centrale de Nantes
sous le sceau de l'Université Bretagne Loire*

École doctorale : Mathématiques et STIC (MathSTIC)

Discipline : Informatique, section CNU 27

Unité de recherche : Laboratoire des Sciences du Numérique de Nantes (LS2N)

Soutenue le 05 décembre 2017

Analyse de réseaux de régulation par approches de coloration de graphes dans le cadre du myélome multiple

JURY

Président : **M. Jean-Daniel ZUCKER**, Directeur de recherche IRD, UMMISCO
Rapporteurs : **M. Mohamed ELATI**, Professeur des universités, Université Lille 1, CRISTAL
M^{me} Nathalie THERET, Directrice de recherche INSERM, IRSET
Examineurs : **M^{me} Carito GUZIOLOWSKI**, Maître de conférence, École centrale de Nantes, LS2N
M^{me} Florence MAGRANGEAS, Docteur en sciences, CHU de Nantes, CRCINA
M. Stéphane MINVIELLE, Directeur de recherche CNRS, CRCINA
Directeur de thèse : **M. Olivier ROUX**, Professeur des universités, École centrale de Nantes, LS2N

Table des matières

1	Introduction	7
1.1	Contexte & Motivation	7
1.2	Données biologiques et évolution des techniques	8
1.2.1	L'expression de gènes	8
1.2.2	Données expression gène : historique	10
1.2.3	Bases de données biologiques : historique	12
1.3	Contributions	15
1.4	Organisation du manuscrit	16
1.5	Notions et méthodes	17
1.5.1	Normalisation de l'expression d'un gène	17
1.5.2	Fouille de données	18
2	État de l'art	23
2.1	Programmation logique	23
2.1.1	Histoire de la programmation logique	23
2.1.2	Answer set Programming	24
2.2	Analyse de l'expression des gènes	26
2.2.1	Introduction	26
2.2.2	Identification de gènes	26
2.2.3	Discrétisation des gènes	27
2.3	Myélome multiple	29
2.3.1	Différenciation normale du plasmocyte	29
2.3.2	Aspect clinique et épidémiologie	30
2.3.3	Évolution et facteurs pronostiques	31
2.4	Des gènes à la fonction : les <i>Pathway Analysis</i>	33
2.4.1	Introduction	33
2.4.2	Over-Representation Analysis (ORA)	33
2.4.3	Functional Class Scoring (FCS)	35
2.4.4	Pathway Topology (PT)	36
2.5	Méthode de coloration cohérente des graphes	36
2.5.1	Introduction	36
2.5.2	Coloration d'un graphe	37
2.5.3	Règles de cohérence des signes	38
2.5.4	Observations	38

2.5.5	Réparations	39
2.5.6	Prédictions et projections	40
3	Modèle de coloration des graphes pour le myélome multiple	41
3.1	Introduction	41
3.2	Traitement des données	41
3.2.1	Discrétisation des données	41
3.2.2	Génération du graphe	43
3.3	Analyse des prédictions	45
3.3.1	Mise en forme des prédictions	45
3.3.2	Validation des prédictions	45
3.3.3	Analyse des prédictions : MM vs NPC	46
3.3.4	Analyse du lien entre survie et prédictions	48
3.3.5	Outils et logiciels	51
3.4	Simulation de l'effet d'un perturbateur	51
3.4.1	Méthode	51
3.4.2	Implémentation	52
3.4.3	Impact des perturbations sur les profils d'expression	52
3.5	Conclusion	52
4	Exploration des colorations parfaites	61
4.1	Introduction	61
4.1.1	Vers la coloration fréquentielle	61
4.1.2	L'énumération des solutions	61
4.1.3	Vers l'identification de sous-solutions par ajout de contraintes	62
4.1.4	Les solutions parfaites	62
4.2	Méthode	63
4.2.1	Modélisation des colorations parfaites en ASP	63
4.2.2	Identification de <i>composants</i>	66
4.2.3	La similarité maximale	67
4.2.4	Réduction de l'espace des solutions	67
4.2.5	Implémentation	69
4.3	Exemple	70
4.3.1	Réduction du graphe	70
4.3.2	Colorations parfaites et identification des <i>composants</i>	71
4.3.3	Calcul de la similarité maximale	72
4.4	Application	73
4.4.1	Données et graphe	73
4.4.2	Colorations parfaites	73
4.4.3	Identification des <i>composants</i>	74
4.4.4	Validation des <i>composants</i>	74
4.4.5	Spécification des <i>composants</i>	75
4.4.6	Analyse biologique des résultats	76

4.5	Comparaison avec d'autres méthodes de classification	78
4.6	Améliorations mises en place	80
4.6.1	Identification des nœuds corrélés en ASP	80
4.6.2	Réduction de l'espace mémoire	81
4.6.3	Amélioration de la réduction du graphe	82
4.7	Conclusion	83
5	Conclusion & perspectives	85
5.1	Contributions	85
5.1.1	La coloration cohérente des graphes	85
5.1.2	Les colorations parfaites	86
5.2	Perspectives	86
5.2.1	Un modèle de perturbations multiples	87
5.2.2	L'intégration de données continues dans les colorations parfaites	88
5.2.3	Perspectives générales	89
	Bibliographie	91
	Annexe	105
.1	Production et communication scientifique	105
.2	Article publié dans Scientific Reports	106
.3	Modèle de perturbations multiples	119

Introduction

1.1 Contexte & Motivation

Depuis une vingtaine d'années on peut observer une explosion de la quantité des données biologiques disponibles. Cette explosion n'a pas épargné le monde de la biologie, et plus précisément celui de la santé qui depuis une dizaine d'années voit la quantité de données et de connaissances augmenter exponentiellement parallèlement avec l'évolution des technologies. Ainsi, le génome humain et ses plus de 3 milliards de paires de bases qui a pu être séquencé en 2003 après 13 ans de travaux et plus de 3 milliards de dollars, peut aujourd'hui l'être en quelques jours pour moins de 1000 dollars. Pour donner un ordre de grandeur, ces 3 milliards de bases correspondent au nombre de caractères de 100 Corans ou 1000 bibles. A cela se sont ajoutées de nombreuses découvertes sur les mécanismes de régulation d'expression des gènes, les interactions possibles avec l'environnement ou même l'héritabilité de certains traits épigénétiques acquis.

Cette augmentation de la quantité de données et de connaissances a permis aux chercheurs d'importantes avancées scientifiques, en particulier en cancérologie directement concernée par les questions de régulation de l'expression des gènes. Un cancer peut se définir par un ensemble de cellules à prolifération incontrôlée formant une ou plusieurs tumeurs. Ces cellules cumulent des modifications leur permettant de les rendre insensibles aux signaux dits de "morts cellulaires" (ou *apoptose*). Aussi la caractérisation des mutations et modifications portées par ces cellules a-t-elle permis d'identifier et de caractériser plus précisément les patients atteints de cette pathologie très hétérogène. Cet objectif d'intégrer le patient dans l'évaluation et les choix cliniques fait partie de l'approche plus générale de la *médecine de précision* (ou médecine personnalisée). A ces fins, se sont développées de nombreuses méthodes visant à identifier les mécanismes sous-jacents aux comportements aberrants de ces cellules cancéreuses. Ces méthodes se sont très rapidement retrouvées à l'interface entre la médecine et la biologie des systèmes, visant à étudier les comportements des organismes en les considérant par le biais des interactions de leurs composants. Ce chevauchement de domaines de recherches a permis, non seulement

d'améliorer la compréhension des mécanismes impliqués dans les cellules cancéreuses, mais aussi de proposer des modèles prédictifs de réactions aux traitements, permettant une meilleure prise en charge des patients.

Cette thèse se place sur cette articulation entre biologie et biologie des systèmes. Nous avons souhaité étudier le myélome multiple par le biais d'approches modélisant la régulation des gènes par coloration de graphes.

Grâce à ces méthodes et aux profils d'expression de patients atteints du myélome multiple, nous avons pu montrer la possibilité d'utiliser ce type d'approches pour inférer l'état des protéines dans un réseau de régulation et s'en servir pour caractériser les voies biologiques mises en jeu dans cette pathologie. Nous avons aussi pu montrer l'intérêt diagnostique de ces approches en comparant ces prédictions aux facteurs cliniques usuellement utilisés. Ensuite, et toujours en se basant sur ces modèles de coloration de graphes, nous avons montré la possibilité de les utiliser pour simuler l'impact d'une perturbation imposée pour un réseau et un profil d'expression donné.

Enfin, nous avons pu proposer une nouvelle approche basée sur la coloration des graphes permettant de discerner plusieurs sous-graphes, et d'associer une métrique de dérégulation pour chacun d'entre-eux à partir d'un profil d'expression. Ces sous-graphes pouvant être associés à des voies biologiques, il nous a été possible d'identifier des fonctions liées à l'apoptose comme dérégulées dans les cellules cancéreuses.

1.2 Données biologiques et évolution des techniques

1.2.1 L'expression de gènes

Un gène est une entité biologique codant pour un ou plusieurs éléments fonctionnels de la cellule. Chez les eucaryotes, les gènes sont transcrits en ARN-pre-messager (ARN-pm), puis après une étape de maturation en ARN-messager (ARN-m). La transcription d'un gène en ARN se fera grâce à une enzyme, l'ARN polymérase qui permettra la synthèse d'un brin d'ARN complémentaire à la séquence d'ADN.

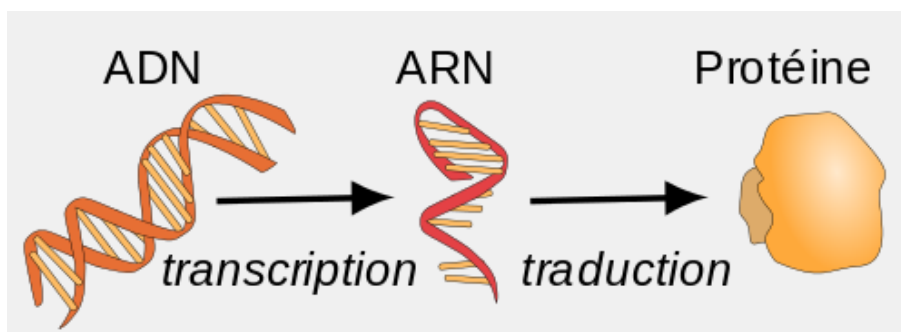


FIGURE 1.1 – Représentation schématique des 2 étapes principales permettant la production de protéines. Tirée de [138]

Après maturation, les ARN-m seront traduits en protéines hors du noyau, au niveau du cytoplasme. Cette traduction mettra en jeu de nombreuses protéines, en particulier les ribosomes,

protéines fortement conservées au cours de l'évolution qui seront chargées de la lecture de l'ARN-m par série de triplets de nucléotides, ceci afin d'associer à chacun d'eux le peptide (brique élémentaire des protéines) correspondant.

D'autres phénomènes ont été mis en évidence plus récemment, montrant le gain en complexité au cours de l'évolution sur ces mécanismes communs à tous les êtres vivants. Lors de la transcription en ARN-pm, celui-ci sera composé d'introns et d'exons. La maturation, aussi appelée "épissage", éliminera les introns et conservera les exons que l'on retrouvera dans l'ARN-m. L'épissage alternatif (*Alternative splicing*) est un mécanisme permettant à partir d'un simple ARN-pm de générer plusieurs ARN-m et donc plusieurs protéines [25]. En effet, lors de la maturation, les exons ne seront pas tous conservés, ni même agencés dans le même ordre. Ce réarrangement permet lors de la traduction d'aboutir à la production de plusieurs protéines (Exemple figure 1.2)

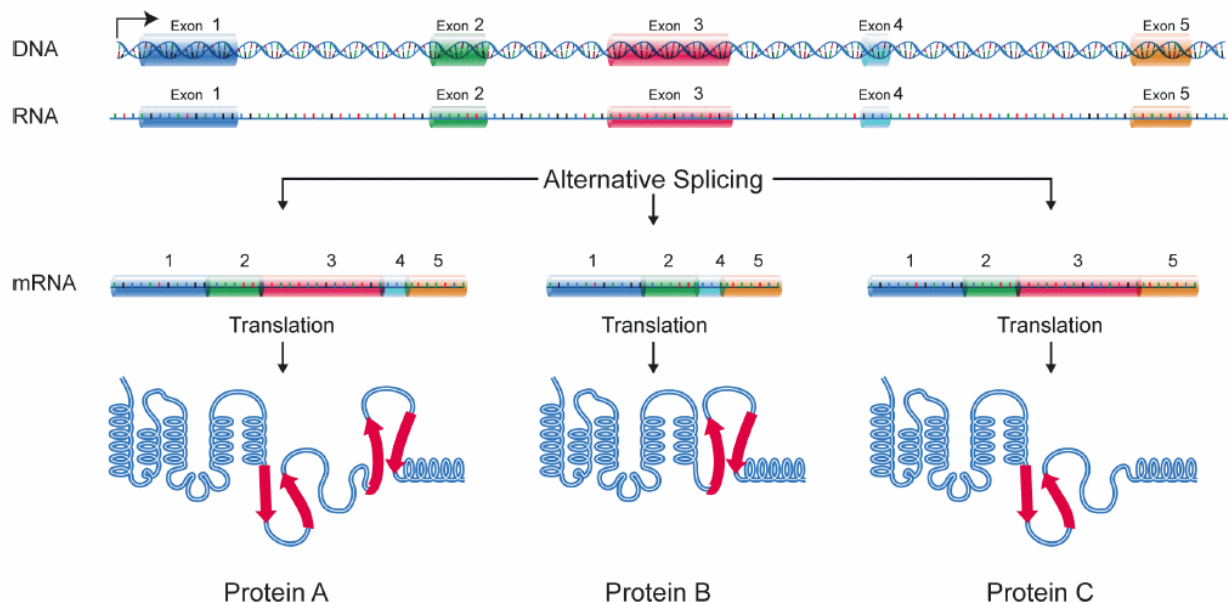


FIGURE 1.2 – Représentation de l'épissage alternatif d'un gène unique conduisant à 3 protéines différentes. Tirée de www.genome.gov

Si la traduction débute normalement sur un triplet AUG (aussi appelé "codon start"), il a été montré que certaines séquences pouvaient elles aussi initier la traduction en protéine. Ces séquences dites "ATIS" (*alternative initiation of translation sites*) semblent présenter de par leur repliement une ressemblance avec un codon-start, entraînant la traduction d'une protéine [130]. Ces séquences, qui semblent conservées au cours de l'évolution [10] sont elles aussi à l'origine de la diversité des protéines.

Enfin, en plus de ces mécanismes, de nombreux facteurs vont pouvoir réguler l'expression des gènes. En effet, si toutes les cellules d'un organisme possèdent le même patrimoine génétique (en dehors des gamètes), l'expression de celles-ci sera extrêmement différente selon le tissu concerné, ou le contexte de la cellule. Parmi ces régulateurs, on trouve les facteurs de transcription, protéines qui interagissent avec l'ADN et l'ARN-polymérase afin d'induire ou au contraire de bloquer la transcription d'un ou plusieurs gènes. Ces facteurs de transcription, eux-

mêmes issus de la transcription puis traduction de gènes, sont extrêmement nombreux [141] et un sujet de recherche très important afin de mieux intégrer les mécanismes de régulation des gènes. Il existe d'autres phénomènes tels que les ARN-interférents (ARN-i) [42], brin d'ARN capables de se fixer sur un brin complémentaire d'ARN-m, amenant à la dégradation du complexe ainsi formé et donc l'inhibition de l'expression du gène associé en empêchant sa traduction. Beaucoup de ces ARN-i proviennent d'introns, ce qui a permis de comprendre un peu plus le rôle régulateur de ces ARN non codants (ne produisant pas directement de protéines). Enfin, certains de ces ARN-i ont été identifiés comme pouvant circuler dans le milieu extra-cellulaire chez l'humain. Ces ARN extra-cellulaires semblent avoir des fonctions biologiques associées à des processus d'exportation très précis bien qu'encore très peu connus [145]. Certains semblent impliqués dans certaines pathologies et ont été caractérisés à des fins diagnostiques, sans pour autant qu'un lien de causalité ait pu être établi [106].

Enfin, ont été mis en avant les phénomènes de régulation dits d'épigénétique. Ces mécanismes s'appliquent sur le génome, et vont, sans modifier sa séquence, rendre accessible ou non des parties de celui-ci pour la transcription. Ce phénomène est un des moteurs principaux de la différenciation cellulaire [51]. Parmi les nombreux processus existants, les 2 plus connus sont le repliement de l'ADN autour des histones et la méthylation de l'ADN. Ces processus semblent intervenir lors de la différenciation cellulaire, ou en réponse à l'environnement. Des corrélations entre certaines pathologies ont aussi été identifiées [31]. Nous pouvons aussi noter que de récentes études tendent à montrer qu'une partie de ces modifications épigénétiques seraient héréditaires, et donc que les descendants pourraient hériter d'une partie de l'histoire vécue et non juste génétique de leurs parents [123].

C'est grâce à tous ces mécanismes, et très certainement bien d'autres encore inconnus, que l'on peut observer une telle diversité dans le vivant. Ainsi si on estime le nombre de gènes entre 20.000 et 25.000 gènes chez l'humain [58], le protéome actuellement connu est d'environ 50.000 protéines et cette quantité augmente encore.

1.2.2 Données expression gène : historique

L'analyse des expressions de gènes est un domaine qui a énormément évolué au cours des 40 dernières années. La mise au point en 1977 du Northern blot, permettant de caractériser l'abondance relative d'une séquence d'ARN, a amené aux premières analyses d'expression de gènes par la mesure de la quantité d'ARN-messager. Cette méthode, simple à mettre en place et peu coûteuse reste encore très utilisée aujourd'hui [104], et a pu voir de nombreuses améliorations concomitantes à l'évolution des connaissances biologiques et à l'automatisation de l'analyse [105]. C'est néanmoins avec l'apparition des premières puces à ADN en 1991 [121] que l'analyse de l'expression des gènes prend une nouvelle ampleur. Celle-ci se base sur la propriété de complémentarité de l'ADN, c'est à dire la faculté qu'un simple brin d'ADN a de s'apparier avec la séquence complémentaire (l'adénine avec la thymine et la guanine avec la cytosine). Une puce consiste en un support (souvent verre ou silicium) sur lequel sera fixé un ensemble de brins courts d'ADN, appelés *sondes* dont les séquences ont été contrôlées durant leurs synthèse. Lors d'une analyse d'un échantillon (figure 1.3), l'ARN-m sera généralement amplifié par PCR (*polymerase chain reaction*) afin d'augmenter la quantité de matériel transcriptomique,

puis converti en ADN-complémentaire (ADN-c) par transcription inverse. Ces ADN-c seront ensuite marqués par une fluorochrome (molécule organique capable d'émettre de la lumière). Enfin, ils seront mis au contact des sondes de la puce afin que les ADN-c se fixent sur celles ayant une séquence complémentaire correspondant au gène étudié. Après nettoyage, seuls les brins d'ADN-c non-fixés seront ôtés de la plaque. Il sera alors possible d'identifier les sondes ayant fixé un ADN-c par l'analyse de la fluorescence de ceux-ci. Une puce peut contenir jusqu'à plusieurs dizaines de milliers de sondes [65], permettant ainsi de couvrir de manière efficace le champ des gènes (et leurs variations) tout en garantissant une qualité d'évaluation de ces niveaux d'expression (en utilisant des sondes redondantes ou spécifiques à des séquences différentes de gènes).

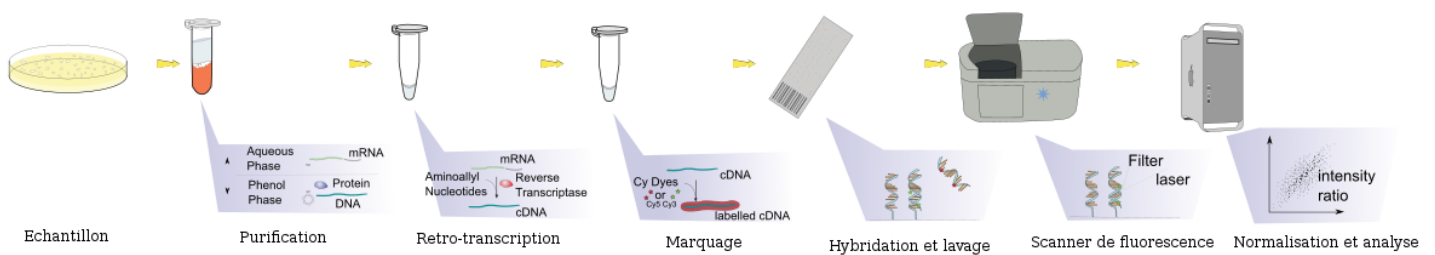


FIGURE 1.3 – Étapes de puce ADN. Image tirée de [140].

Ainsi, ces puces permettent d'analyser de grandes quantités de gènes en même temps permettant ainsi leur utilisation pour comparer des profils d'expression entre eux [76, 107, 33]. Ces puces vont se développer au cours des 2 décennies, avec une augmentation forte de leurs capacités corrélée à une réduction du coût et du temps nécessaire pour ce type d'analyse. Néanmoins, celles-ci présentent quelques limites. En effet, l'utilisation des sondes nécessite de connaître préalablement les séquences à étudier et donc les gènes associés. Aussi, ce type d'analyses, bien que très efficaces ne permettent de travailler qu'à partir d'une connaissance à priori des gènes et de leur séquences et ne peut amener à identifier de nouveaux gènes. Cependant, dans le cas de l'humain, cette problématique se pose moins depuis la fin du projet "génomique humaine" qui a permis de séquencer l'intégralité du génome humain en une dizaine d'années [119].

Une nouvelle révolution dans l'analyse de l'expression des gènes aura lieu vers 2008 [99] avec les premiers séquenceurs à haut débit, ou NGS (*next-generation sequencing*). Celles-ci se basent sur les approches de séquençage haut débit de l'ADN, utilisable après retro-transcription des brins d'ARN-m en ADN-c. Auparavant, les séquençages étaient le plus souvent effectués via la méthode dites de Sanger [118], méthode permettant de séquencer un brin à la fois via 4 amplifications parallèles par PCR contrôlée. Ces amplifications (une par nucléotide) amenaient à obtenir pour un brin d'ADN initial une série de brins complémentaires de taille variable et sur lesquels, le dernier nucléotide modifié (un didésoxyribonucléotide, dont la fixation empêche la poursuite de la synthèse) était connu. Par électrophorèse (migration sur gel), il était alors possible de distinguer les brins sur la base de leurs poids et donc de connaître la position de chaque nucléotide (Exemple figure 1.4).

Néanmoins, cette méthode restait limitée car ne pouvant analyser qu'un fragment à la fois malgré d'importantes améliorations dont l'automatisation du procédé et l'ajout de fluorophores

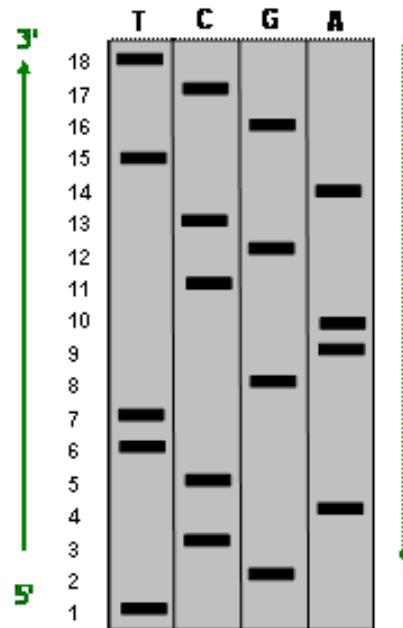


FIGURE 1.4 – Exemple de résultat d'électrophorèse avec la méthode de Sanger. La séquence correspondante sera : 5' - TGC ACTTGAACGCATGCT - 3'

sur les didésoxyribonucléotides de manière à n'avoir qu'une seule PCR à effectuer. Les séquenceurs à haut débit offrant la possibilité de séquencer l'intégralité des fragments d'ADN présents dans un échantillon, ceux-ci ont permis de séquencer des génomes et transcriptomes entiers, faisant entrer la recherche dans une ère que certains n'hésitent pas à qualifier de méta-omique, en allusion à la racine latine [74]. Cette nouvelle ère de recherche pose néanmoins de nouveaux problèmes dans le monde de la recherche. L'augmentation de la quantité des données (on parlera de méta-omique horizontale) a nécessité de développer de nouvelles méthodes permettant de les stocker, traiter et analyser amenant au développement du champs de la bio-informatique analysant les séquences. De la même manière, si le partage des informations et des connaissances entre chercheurs a été facilité avec le développement de bases de données centralisant celles-ci [48], ces données sont devenues bien plus complexes à comprendre car issues d'une série de processus techniques de plus en plus poussés et multidisciplinaires [102]. Ces problématiques risquent de s'accroître non seulement avec l'accélération des capacités de production de données (figure 1.5) mais aussi avec l'arrivée plus récente des approches multi-omiques (on parlera ici de méta-omique verticale), proposant de combiner l'information à plusieurs échelles (génomique, transcriptomique, protéomique, épigénomique, population, microbiome, etc.)

1.2.3 Bases de données biologiques : historique

Avec cette augmentation exponentielle de la quantité des données et connaissances biologiques, ainsi que de leur hétérogénéité, de nombreuses bases de données ont été mises en place intégrant celles-ci afin de tenter de répondre au besoin de les structurer pour faciliter leur partage. Celles-ci ont vu leur nombre augmenter drastiquement, ainsi la base de données du journal *Nucleic Acids Research* qui référençait 96 bases dans sa première version en 2001 [9] en conte-

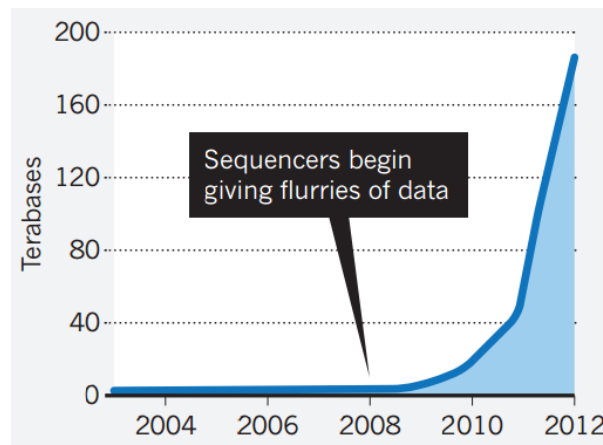


FIGURE 1.5 – Évolution de la quantité de données génétiques stockées à l'institut européen de Bioinformatique (EBI). Tiré de [88]

nait 1380 en 2012 [43]. En miroir à l'hétérogénéité des données et connaissances biologiques, ces bases de données sont très variées bien que quelques unes tentent d'être plus généralistes telles que la base *Ensembl* [3]. Certaines vont se spécialiser sur certains types d'organismes, par exemple la base *Saccharomyces Genome Database* (SGD) [27]. D'autres vont plutôt s'intéresser uniquement à l'humain et s'orienter sur des spécificités vis à vis de certaines pathologies. Parmi celles-ci, 2 des plus connues sont l'*International Cancer Genome Consortium* (ICGC) [30] et *the Cancer Genome Atlas* (TCGA) [136], toutes les deux spécialisées dans la cancérologie et proposant des données génomiques, transcriptomiques, épigénomiques et cliniques pour plusieurs types de cancers. Enfin, d'autres bases de données se sont tournées non pas sur le stockage de données biologiques mais sur celui des connaissances. Parmi ces bases, l'une des plus connues Uniprot [128] qui se divise en 4 sous-bases permet le partage d'information sur les protéines, aussi bien leurs séquences, que les gènes, fonctions et organismes associées. Cette base est actuellement une des références en analyses protéomiques, l'identifiant dit "Uniprot" (clé primaire de cette base) est couramment utilisé. D'autres bases se focaliseront plus sur les informations et leur implication dans les fonctions biologiques des gènes comme la base *Gene ontology* (GO) [127]. D'autres bases se sont plus focalisées sur les voies biologiques, amenant à considérer gènes, protéines, molécules, etc. sur la base de leurs interactions avec les autres composants des systèmes biologiques. Une des plus connues, *Kyoto Encyclopedia of Genes and Genomes* (KEGG) [64] permet d'avoir accès à un ensemble de voies biologiques (*pathways*) ainsi que les acteurs dans chacune de celles-ci (voir exemple figure 1.6). D'autres bases, plus spécialisées comme *Pathway Interaction Database* (PID) [120] ou *Causal-Bionet* [19] proposent elles-aussi des informations sur les *pathways* impliqués en cancérologie. Enfin la base de données *TRRUST* [56] propose de stocker les informations sur les réseaux de régulation transcriptionnel. Son originalité tient du mode de traitement des informations, qui sont identifiées par *text mining* dans les résumés de publications puis corrigées manuellement. Cette liste bien que très loin d'être exhaustive, permet d'apprécier la diversité de ces bases de données.

Cette diversité n'est pas sans poser de nombreuses interrogations et difficultés pour les chercheurs [47]. Certaines de ces bases présentent des limites conceptuelles, empêchant une utili-

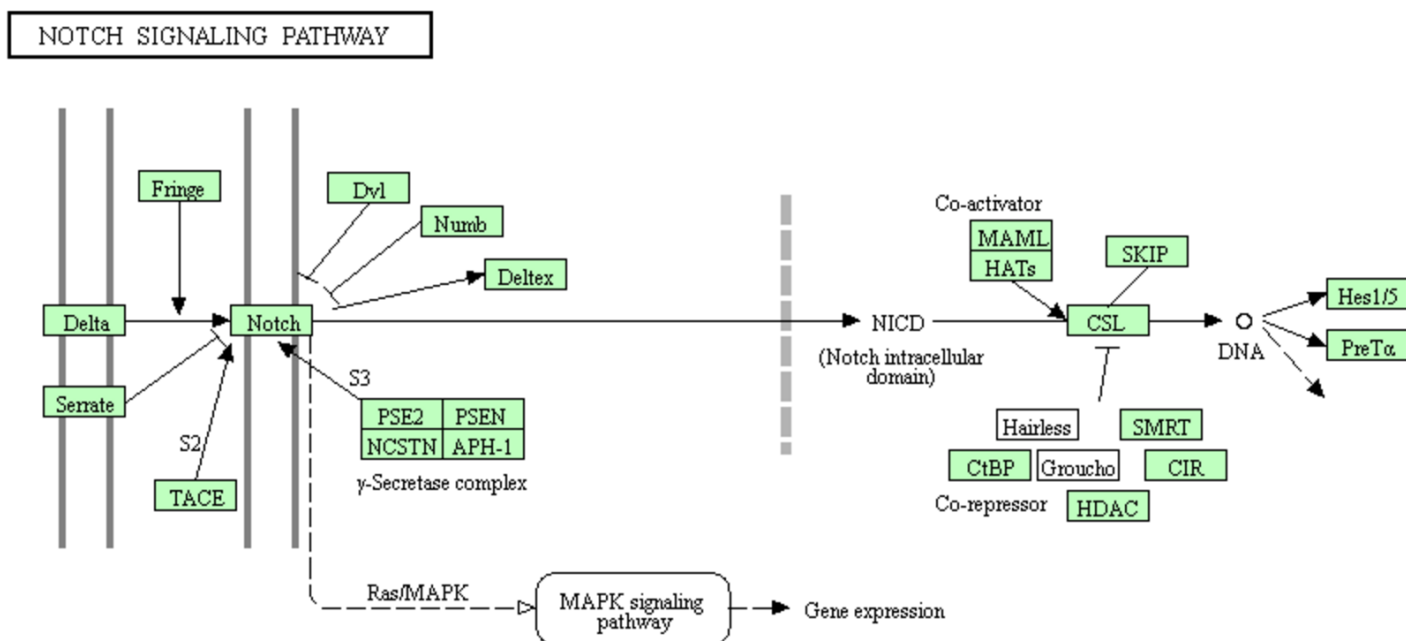


FIGURE 1.6 – Exemple de représentation graphique de la voie de signalisation Notch sur KEGG

sation par des approches "automatisées", ainsi la base *Causal BioNet* (CBN) n'est accessible que par l'interface graphique du site associé. D'autres bases se sont vues mises en ligne à l'état d'ébauche, aussi les nombreuses mises à jour ont empêché la stabilité de fonctionnement pouvant permettre le développement et l'utilisation de logiciels les utilisant. De nombreuses bases ont développé leur propre sémantique, en particulier sur l'identification des éléments biologiques. Aussi les chercheurs sont donc obligés de passer par une étape de traduction des divers termes vers une terminologie commune. Très souvent, cette traduction passe par une perte d'information. Ainsi PID stocke les *pathways* au format XML et contient des informations de modifications protéiques, elles mêmes distinguées par leur identifiant Uniprot tandis que CBN conserve les *pathways* au format *Biological Expression Language* (BEL) avec les protéines (sans informations de modification) au format *HUGO Gene Nomenclature Committee* [49]. Enfin une autre problématique forte se pose vis à vis de la qualité des données. Si certaines bases sont vérifiées manuellement d'autres le sont par des approches automatisées permettant en contrepartie d'incorporer plus d'information. Cette balance difficile à équilibrer a aussi amené à des entre-deux comme Uniprot, qui contient 2 sous-bases [16] : Swiss-Prot corrigée manuellement et Trembl dont la correction est automatique.

Nous pouvons néanmoins signaler la tendance plus récente à la fusion de ces bases de données en bases uniques, tout comme l'avait fait la base Uniprot. Pour de nombreuses raisons (financière, passage de l'effet de mode, valorisation faible), de nombreuses bases ont fermé ou ont cessé d'incorporer de nouvelles données, tandis que le rythme de création de nouvelles bases semble se ralentir. Par exemple, la base de données du journal *Nucleic Acids Research* précédemment citée n'a, lors de sa mise à jour 2017, vu qu'une augmentation de 22 bases (70 ajouts et 48 suppressions) [44]. Ainsi certaines nouvelles "méga-bases" proposent-elles aujourd'hui de mutualiser toutes ces données en un interface unique. Ainsi la base *NDex* [108] stocke de

nombreux *Pathways* au format de leur base initial. Celle-ci présente en plus l'avantage de permettre le dépôt de ses propres *pathways* et de contrôler leur diffusion. Néanmoins cette base ne passe pas par une uniformisation des données, la limitant à une fonction de dépositaire. D'autres bases, elles, présentent l'avantage d'offrir une sémantique qui se veut la plus généraliste possible. Ainsi la base *Pathway commons* [23] conserve l'intégralité de 25 (21 initialement) bases de données au format Biopax. Enfin, la base *wikipathways* [72] se veut à mi-chemin entre les 2 méga-bases citées préalablement. Basée sur le même principe que Wikipédia, elle présente d'un côté une sémantique unique, et de l'autre la possibilité à chaque utilisateur de déposer et/ou corriger des *pathways*. A cela s'ajoute son portail avec des bases de données telles que Reactome permettant aussi de profiter des données stockées sur ces bases. Enfin, il est important de signaler une base plus récente, *Omnipath* [131] visant à réunir les *pathways* spécifiques à l'humain en se basant sur 34 bases de données stockant des interactions à partir de la littérature. Cette base couvre actuellement 39% du protéome humain.

1.3 Contributions

Cette thèse présente 2 contributions majeures qui ont pu être effectuées durant ce doctorat.

La première contribution a été la première modélisation de données transcriptomiques issues de 602 patients atteints du myélome multiple et de 9 patients sains par des méthodes de colorations cohérentes de graphes en utilisant un réseau de régulation extrait de la base PID. Par cette approche, il a été possible d'inférer l'état des éléments non-observés de ce réseau de régulation et de les utiliser afin de caractériser chaque profil d'expression. Nous avons ensuite proposé une méthode de comparaison des colorations, permettant de comparer les profils entre eux. Ces prédictions d'état ont permis d'identifier des bio-marqueurs spécifiques dans le myélome multiple et déjà connus dans la littérature, confirmant ainsi la méthode. Nous avons ensuite pu caractériser l'activité de protéines et de fonctions biologiques comme pouvant servir de facteur pronostique enrichissant les modèles prédictifs basés sur les données bio-cliniques. Enfin, nous avons proposé un modèle plus exploratoire, basé sur la coloration cohérente des graphes afin de simuler l'impact d'une perturbation unique sur un réseau et un profil d'expression de gènes. Ces résultats ont été publiés en 2017 dans le journal *Scientific Reports* [93]

La seconde contribution a porté sur la proposition et l'application d'une nouvelle méthode d'identification de sous-réseaux par colorations dites *parfaites* de graphes. A partir d'un réseau modélisé sous forme de graphe, cette approche identifie les ensembles de nœuds dont les signes (ou colorations) seront corrélés entre eux et les assemble en sous-graphes appelés *composants*. De plus, à partir d'un profil d'expression de gènes, il est possible de calculer l'écart entre ce profil et les états *parfaits* de ces sous-réseaux que nous avons appelé la *similarité maximale*. Nous avons appliqué cette méthode sur le même réseau que précédemment afin d'identifier les *composants*. Le calcul de la *similarité maximale* sur les données des 611 individus a permis d'identifier un de ces *composants* comme spécifiquement dérégulé chez les profils atteints du myélome multiple. Ce même *composant* a pu être ensuite associé à des fonctions biologiques impliquées dans les processus oncogéniques. Ces résultats ont été présentés lors du *workshop* CNB-MAC [94] qui a eu lieu le 20 août 2017 à Boston et sont actuellement en révision mineure

pour une publication journal.

Enfin, une dernière contribution, non-présentée dans cette thèse aura été la participation au DREAM-CHALLENGE 9, portant sur la prédiction de la résistance au traitement de patients atteints de leucémie myéloïde aiguë (*Acute Myeloid Leukemia* : AML). L'objectif était de distinguer 2 classes de patients, les répondants et les résistants, en se basant sur des mesures de phospho-protéomiques. Nous avons travaillé avec le modèle nommé caspo permettant à partir d'un réseau initial (*Prior Knowledge Network* : PKN) et d'un ensemble d'observations d'identifier les familles de réseaux booléens expliquant le mieux ces observations. Il est alors possible de générer 2 familles de réseaux booléens, correspondant aux patients répondants et résistants et de calculer la distance avec un nouveau jeu d'observations. La problématique posée par cette méthode était la pré-sélection des protéines afin de générer un PKN de taille suffisamment petit pour l'apprentissage des réseaux booléens associés. Nos résultats préliminaires n'ont pas permis d'établir un modèle suffisamment satisfaisant, avec un taux de réponse estimé à 37.5% (pourcentage de patients ayant une distance différente entre les 2 familles de réseaux booléens) et un taux de précision de 22% (ou 58.7% en considérant uniquement les réponses) et ceci, sur les données d'apprentissage. Néanmoins, ces travaux préliminaires ont permis la poursuite sur la sélection des groupes de protéines à utiliser, amenant à un nouveau modèle ayant une précision de 69%. Ces travaux ont été soumis dans le journal BMC Bioinformatics.

1.4 Organisation du manuscrit

Ce manuscrit de thèse s'organise en 5 chapitres. Le premier chapitre, en plus de son rôle introductif et de contextualisation présentera aussi des notions et méthodes qui paraissent non-essentiels pour la compréhension du reste du manuscrit. Néanmoins, du fait du caractère multidisciplinaire de cette thèse, il semblait important de prendre en compte la diversité des profils des potentiels lecteurs. Aussi, certaines notions pourront paraître triviales et leur présentation superficielle pour certains.

Le second chapitre visera à présenter plus en détail le contexte de connaissances dans lesquelles nous avons mené ces travaux. Dans un premier temps, sera présenté l'aspect plus technique, à savoir la programmation logique et les méthodes d'analyses de gènes. Ensuite, seront présentées les connaissances actuelles sur le myélome multiple. Enfin, ces travaux s'inscrivent dans la lignée des méthodes de *pathways analysis*, un tour d'horizon des méthodes principales de ce domaine. Puis, nous nous focaliserons sur la coloration cohérente des graphes et leur fonctionnement.

Le troisième chapitre présentera la première contribution originale de cette thèse, à savoir l'application de la méthode de coloration cohérente de graphe au myélome multiple et de manière plus générale la tentative de l'utiliser dans une logique de *pathways analysis*, ainsi que la modélisation proposée des perturbations simples sur un réseau et un profil d'expression.

Le quatrième chapitre aura pour objectif de présenter la méthode d'identification des sous-réseaux appelés *composants* sur la base de la recherche des colorations *parfaites* de graphes. Nous montrerons ensuite comment l'intégration de données d'expression est possible en estimant la similarité entre les configurations de ces *composants* et ces profils d'expression. En troi-

sième lieu, nous présenterons les réductions de graphes mises au point afin de réduire le temps de calcul. Puis, seront présentées 2 applications de cette méthode de colorations parfaites. La première portera sur un exemple simple afin d'illustrer le fonctionnement de la méthode tandis que la seconde utilisera les données utilisées dans le chapitre précédent afin de montrer le potentiel intérêt de cette méthode sur des questions biologiques. Enfin, nous présenterons des améliorations de cette méthode plus récentes permettant de réduire le temps de calcul pour identifier les *composants*.

Pour terminer ; le chapitre 5 proposera une rétrospective des apports de cette thèse ainsi que les perspectives potentielles. Nous profiterons de ce chapitre pour présenter 2 modèles théoriques, qui pourraient constituer des axes de réflexion pour une poursuite de ces travaux.

1.5 Notions et méthodes

1.5.1 Normalisation de l'expression d'un gène

L'expression d'un gène est mesurée par un ratio de fluorescence liée à la "quantité" d'ARN lui correspondant dans un échantillon. De par ce fait, la distribution de cette expression ne sera pas normale (distribution symétrique, ayant la même valeur pour la moyenne, le mode et la médiane). Aussi, en l'état n'est-il pas possible d'utiliser la plupart des outils statistiques classiques, ayant pour pré-requis, cette normalité de distribution. Afin de pouvoir analyser par des approches statistiques ces expressions de gènes, il est nécessaire de passer par une phase de transformation de celles-ci. Le logarithme en base 2 sur les données d'expression est la méthode la plus souvent utilisée [110] afin de normaliser leurs distributions (figure 1.7).

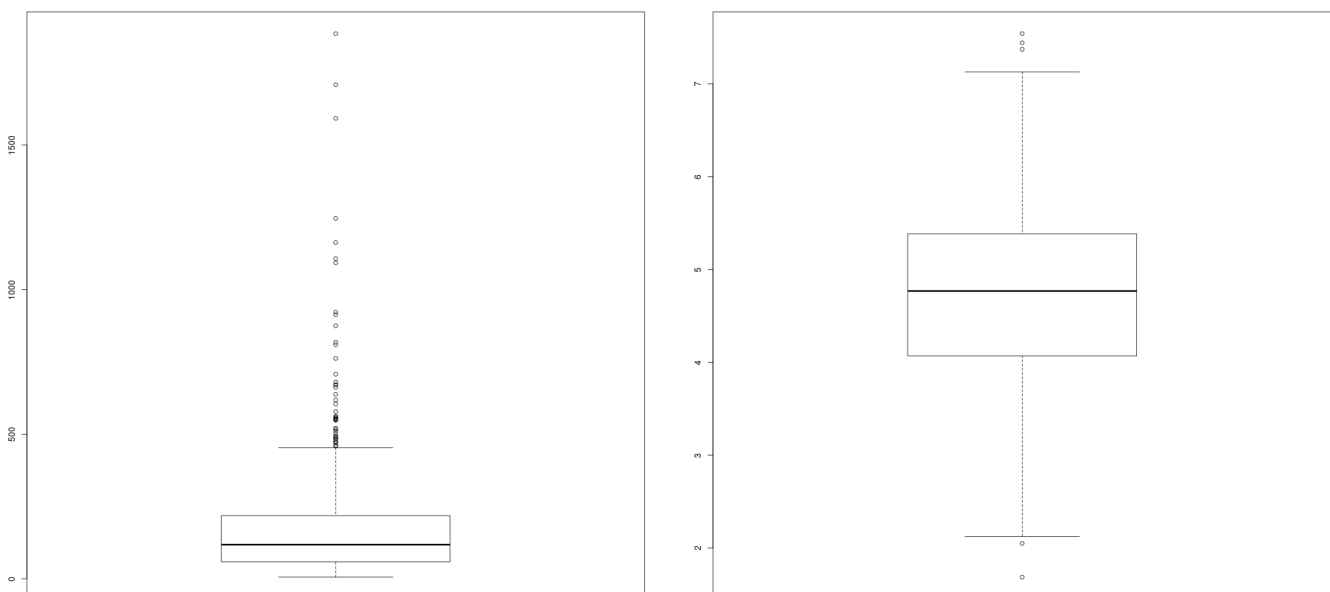


FIGURE 1.7 – Exemple de normalisation de données d'expression d'un gène. A gauche : répartition de la mesure de fluorescence. A droite : même distribution après une normalisation par logarithme 2.

Il est ainsi possible sur ces données transformées d'appliquer les outils statistiques classiques permettant, entre autre, d'identifier des gènes variants. Dans ce manuscrit, nous travaillerons uniquement avec des données transformées par logarithme 2. Pour des raisons de facilité de lecture, nous parlerons de données d'expressions de gènes pour ces distribution normalisées.

1.5.2 Fouille de données

L'analyse de ces données a nécessité le recours à des approches automatisées dites de "fouilles de données" (*data mining*), méthodes permettant d'extraire des informations à partir d'une grande masse de données. Ce domaine de recherche n'est pas spécifique à la bio-informatique, qui n'en représente qu'une petite fraction (14% des données traitées sur <http://www.kdnuggets.com/>). On peut les diviser en 2 catégories : les méthodes de classification non-supervisées (*clustering*) et les méthodes de classification supervisées. Dans les sous-sections suivantes, nous présenterons 3 méthodes utilisées dans cette thèse appartenant à ces catégories.

Les approches non-supervisées permettent de structurer un ensemble de données (n observations de v variables) sans informations autre que ces données. Ainsi, il est possible de diviser cet ensemble de données en plusieurs *classes* selon des critères prédéfinis (nombre, taille, etc.). Ce type d'approches nécessite d'établir préalablement la méthode de calcul de distance entre les variables. Cette distance peut être basée sur la distance euclidienne, la distance de Hamming (pour les données discrètes), ou des distances pondérées selon les variables. Enfin, un second choix est à établir préalablement, la métrique de distance entre les clusters (rassemblement d'observations). Il est possible de considérer la distance minimale, moyenne, maximale, médiane entre les clusters comme critère pour les assemblages de ceux-ci. Nous présenterons la méthode des k -moyennes appartenant à ce type d'approches.

Les méthodes supervisées, elles, vont fonctionner en 2 phases. La première, dite phase d'apprentissage va utiliser un jeu de données dont les classes sont connues (ensemble d'apprentissage) afin de générer un modèle de classification. La seconde phase, dite phase de test, va chercher à attribuer une classe (ou une probabilité d'appartenance à chaque classe) pour un ensemble de données dont les classes sont inconnues. Nous présenterons ici 2 méthodes utilisées dans ce manuscrit, les arbres de décision et les forêts aléatoires.

1.5.2.1 Méthode des K -moyennes

La méthode de partitionnement en k -moyennes (*K-means*) a été proposée en 1956 par H. Steinhaus [125], son algorithme sera publié en 1982 par S. Lloyd [80].

Cette méthode de classification non-supervisée consiste, pour un jeu de n observations et v variables à répartir en k clusters, à classer chaque observation sur la base d'une minimisation de l'hétérogénéité à l'intérieur des clusters (Exemple d'application figure 1.8). Pour cela, chaque observation va être modélisée par un point dans un espace à v dimensions. k points seront placés, appelés *centres initiaux*. Ces placements peuvent soit être aléatoires, soit être définis au départ. Pour chaque observation sera calculée son appartenance à une de ces classes sur la base de la distance avec ces centres initiaux. A partir de ces clusters établis, il est possible alors de recalculer un nouveau centre pour chaque classe en se basant sur les valeurs moyennes. L'algorithme recalculera alors l'appartenance de classe pour chaque observation puis les nouveaux centres

jusqu'à ce que la condition d'arrêt soit satisfaite. Cette condition d'arrêt peut être de 2 types. La plupart du temps, l'algorithme s'arrêtera quand le calcul des nouveaux centres n'amène plus de changement dans la distribution des observations dans les classes (stabilité de la répartition). Il est aussi possible de fixer un nombre maximal d'itérations. Cette condition peut être utile en particulier sur des distributions très condensées.

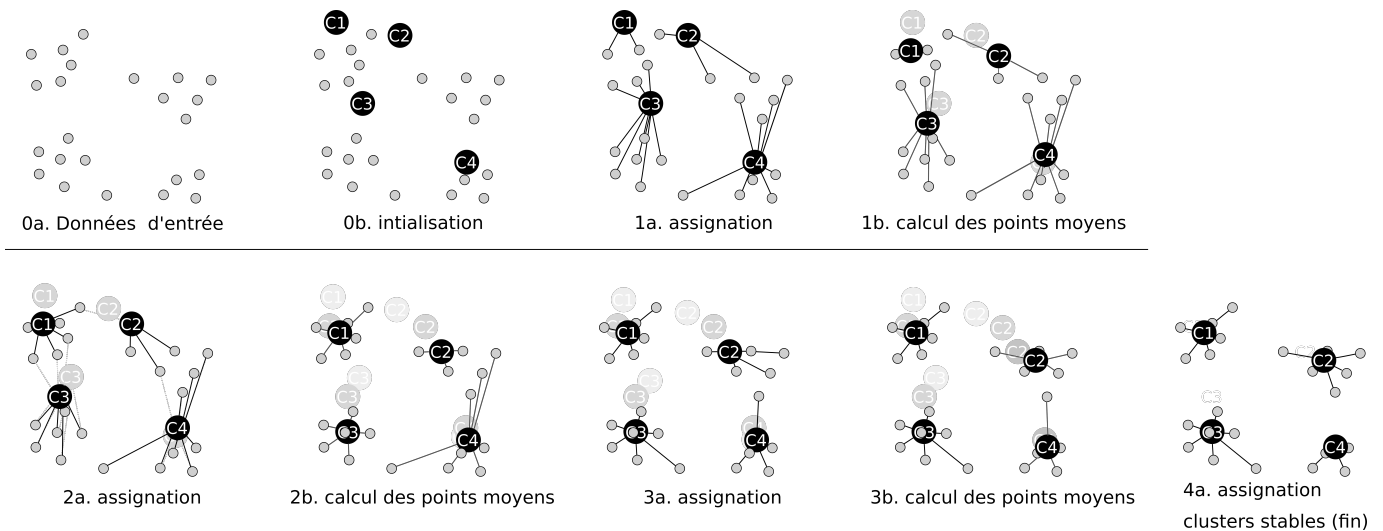


FIGURE 1.8 – Illustration du fonctionnement de la méthode des K-moyennes avec 4 classes. Tirée de [139].

Il est à noter que le calcul des nouveaux centres de gravité peut aussi se baser sur les valeurs médianes, on parlera alors de K-médianes. Cette méthode, bien qu'efficace en pratique pour partitionner des données, présente quelques limites et dépendances. Le nombre de classes et le placement des centres initiaux influenceront énormément le résultat. De plus, on peut obtenir dans certains cas, des répartitions non-stables, d'où l'intérêt d'un nombre maximal d'itérations.

1.5.2.2 Arbre de décision :

La méthode par arbres de décision (*decision tree*) [96] est une approche de classification supervisée permettant d'établir les combinaisons de variables spécifiques à chaque classe prédéterminée. L'algorithme associé est itératif, et va chercher pour chaque étape à identifier la variable la plus discriminante entre les classes, puis séparer les observations en 2 groupes (selon la variable et la valeur de celle-ci) qui seront à leur tour analysés pour identifier la variable la plus discriminante entre les 2 sous-échantillons. Les conditions d'arrêts peuvent être la profondeur de l'arbre (le nombre de variables à utiliser) ou la précision (la "pureté" de chaque sous-échantillon). Cette méthode présente l'avantage d'être assez simple à mettre en place et à appliquer sur un jeu de données. De plus, le modèle obtenu est facilement représentable graphiquement C'est l'une des raisons pour laquelle cette méthode est utilisée dans de nombreux domaines dont l'aide à la décision . Nous présentons ici un exemple d'arbre de décision pour du diagnostic clinique à partir des données du tableau 1.1. La classe à prédire est celle de la colonne "Maladie". Les autres variables seront donc utilisées afin de déterminer pour un individu

la classe (et donc la maladie) à lui attribuer. Le résultat de cet apprentissage est représenté dans la figure 1.9

TABLE 1.1 – Données en entrée pour l'apprentissage de l'arbre de décision de la figure 1.9.

Fievre	Douleur	Toux	Maladie
oui	Abdomen	non	Appendicite
non	Abdomen	oui	Appendicite
oui	gorge	non	rhume
oui	gorge	oui	rhume
non	gorge	oui	mal de gorge
oui	non	non	aucune
oui	non	oui	rhume
non	non	oui	refroidissement
non	non	non	aucune

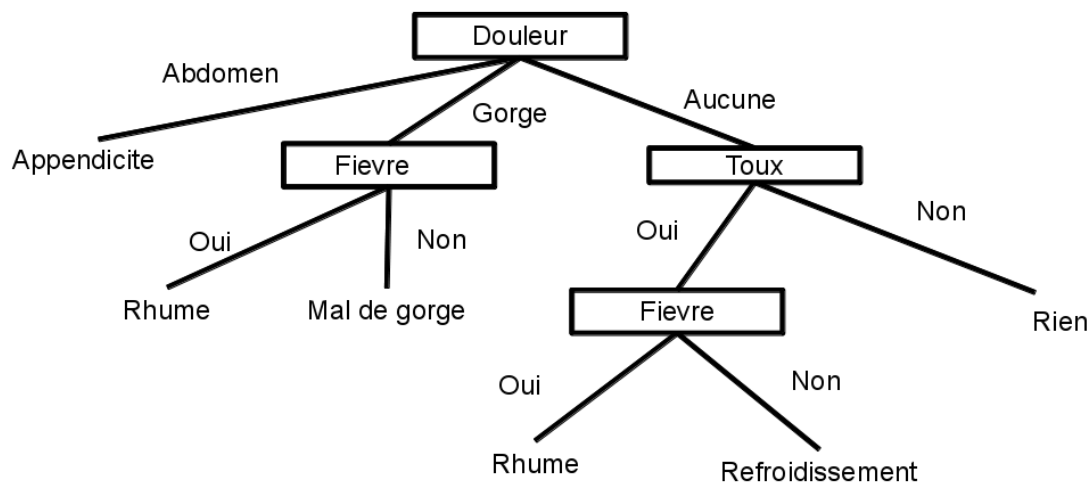


FIGURE 1.9 – Arbre de décision généré à partir des données cliniques. Les nœuds sont les variables utilisées à chaque étape. Les feuilles (nœuds terminaux), sont les classes associées.

Une manière classique de tester la qualité d'un arbre consiste à le tester avec un jeu de données dont on connaît les classes d'appartenance et de comparer celles-ci avec les classes prédites. Ainsi, on peut évaluer la précision du modèle de décision. Néanmoins, les choix sur la condition d'arrêt peuvent amener à un sur-apprentissage (modèle trop adapté aux données d'apprentissage et incapable de prédire sur un autre jeu de données). De plus, cette méthode est très sensible aux données aberrantes. De la même manière, la répartition des effectifs entre les classes est à prendre en compte. Enfin, le choix des variables discriminantes est un choix de minimisation locale de l'inertie, ce qui peut amener à un modèle prédictif non-optimal au niveau global.

1.5.2.3 Forêts aléatoires

L'approche par forêt aléatoire (*random forest*) [38] est une amélioration des arbres de décision visant à pallier certains de ses inconvénients en générant non plus un arbre, mais un ensemble d'arbres (d'où le terme de forêt) de profondeur réduite. Pour un jeu de n observations et v variables à répartir en k clusters, chaque arbre sera appris à partir d'un sous-ensemble aléatoire d'observations et de variables, de telle manière à ce que chacun soit le plus différent possible des autres. Ainsi pour chaque arbre généré, il est possible d'obtenir une classe (ou une probabilité d'appartenance). Le choix final étant généralement le choix (ou la probabilité) majoritaire. Bien que cette méthode fasse perdre la représentation graphique d'un arbre de décision, il reste possible d'évaluer l'information apportée par chaque variable en identifiant leur fréquence de présence dans ces arbres (pondérée par la hauteur). Voir un exemple de représentation du poids des variables dans la figure 1.10.

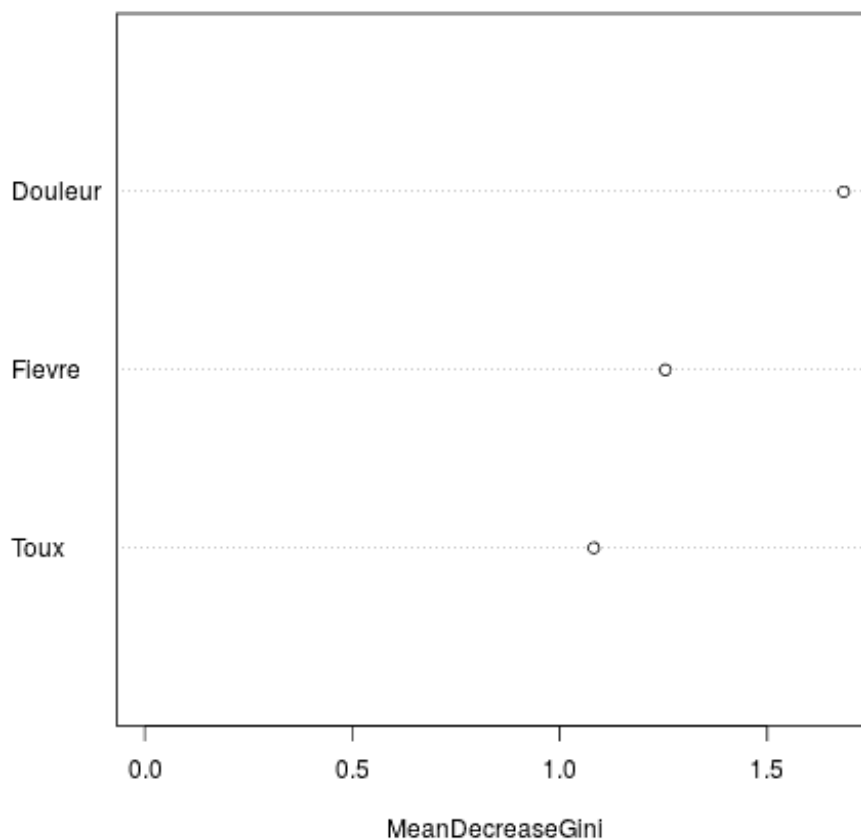


FIGURE 1.10 – Poids des variables sur les données présentées dans le tableau 1.1 dans un modèle prédisant la maladie.

On notera aussi que l'apprentissage du modèle peut être très long sur des données de grandes dimensions (car nécessitant de créer un ensemble important d'arbres), néanmoins la prédiction sur de nouvelles données reste rapide. Enfin, cette méthode est bien moins sensible aux données extrêmes grâce aux échantillonnages réguliers sensés cacher ce bruitage rendant cette méthode utilisable pour des analyses transcriptomiques [2].

État de l'art

2.1 Programmation logique

Dans cette partie, nous présenterons brièvement le développement de la programmation logique. Nous nous focaliserons sur l'Answer Set Programming (ASP), langage utilisé dans les approches montrées dans les chapitres suivants. Nous présenterons aussi quelques fonctionnalités utilisées dans ces approches.

2.1.1 Histoire de la programmation logique

La programmation logique est une forme de programmation déclarative très utilisée dans les champs de l'intelligence artificielle basée sur un ensemble de phrases logiques, exprimant des faits et des règles sur un problème à résoudre.

Si les conditions de validité d'une démonstration automatique ont été théorisées dès 1930 par Herbrand [57], il faudra attendre 1969 pour voir apparaître un premier langage de programmation logique, Absys [37]. Ce langage ne connaîtra pas le même succès que son successeur qui sortira en 1972, Prolog [28] et aura de nombreuses améliorations au cours de son développement. Le Prolog se base sur 3 concepts majeurs :

- L'unification : rendre identique des termes en remplaçant les variables
- La récursivité : un processus qui appellera le même processus.
- Le retour sur trace (ou *backtracking*) : l'exploration des solutions se fait de manière incrémentale. La construction d'une solution partielle (ou complète) est abandonnée dès qu'il est identifié que les solutions qui en découleront ne seront pas satisfaisantes.

Ces 3 concepts-clés rendent le Prolog tout à fait adapté pour les recherches exhaustives de solutions dans un problème logique. En 2003, l'Answer Set Programming [8] est présenté afin

de permettre la résolution de problèmes combinatoires. Tout comme le Prolog, l'ASP est un langage logique, où le problème doit être décrit par l'utilisateur avant que le solveur ne cherche les solutions respectant les faits, règles et buts décrits. Les solutions identifiées par le solveur sont appelées "ensembles réponses" (ou *answer set*). La syntaxe est aussi très proche du Prolog. En revanche, contrairement au Prolog, l'ASP [78, 45] inclut l'opérateur de négation, ainsi que la possibilité d'exprimer les critères à optimiser. De plus, en ASP, l'ordre d'énonciation des phrases n'a pas d'incidence.

2.1.2 Answer set Programming

2.1.2.1 Syntaxe et prédicats

Un programme logique en ASP contient un ensemble de règles, de faits et de contraintes. Le programme logique écrit sera traité par un solveur qui cherchera les modèles stables (les *answer sets*) répondant aux contraintes, faits et optimisations fixés dans le programme logique.

Les règles sont composées de 2 éléments-clés, la tête (à gauche) et la queue (à droite) séparées par le signe \leftarrow , eux-mêmes composés d'un ou plusieurs atomes (élément non-décomposable d'un programme). Le symbole \leftarrow peut être traduit par l'instruction algorithmique "SI". Notons aussi que l'opérateur de la négativité est **not**. Enfin, un prédicat est terminé par un point.

```
1 A0  $\leftarrow$  A1, ..., An, not An+1, ..., not An+k.
```

Ainsi la ligne 1 explicite le fait que A_0 sera *vrai* **SI** A_1, \dots, A_n sont **vrais** et A_{n+1}, \dots, A_{n+k} **ne sont pas vrais** (ne sont pas présents dans la solution explorée). Un *answer set* sera un ensemble d'atomes tel que toutes les règles logiques sont satisfaites. Notons qu'une variable commence avec une majuscule et qu'une constante commence avec une minuscule.

Un fait est une règle sans queue et n'a donc pas à être vérifié. L'exemple des lignes 2 à 5 présente un cas d'instanciation d'un problème simple. 3 faits sont définis : le chien et la pie sont des animaux (Ligne 2). La pie est un oiseau (Ligne 3). Ces faits n'ayant pas de conditions, ils sont considérés comme vrais et serviront à déduire de nouveaux prédicats. Enfin, une règle est définie : les oiseaux ont des plumes (Ligne 4). La dernière ligne (Ligne 5), permet d'indiquer que nous ne souhaitons afficher que les prédicats de type "plume".

```
2 animal(chien;pie).
3 oiseau(pie).
4 plume(X)  $\leftarrow$  oiseau(X).
5 #show plume/1.
```

2.1.2.2 Génération d'answer sets

Il est possible en ASP de créer des ensembles d'atomes sans avoir à définir toutes les valeurs en utilisant la structure présentée à la ligne 6. Dans ce cas-là, entre n et m prédicats de type $a(X, Z)$, où X sera sélectionné à partir du domaine des prédicats décrits par $b(X)$, seront générés pour chaque *answer set* et pour chaque atome vrai $c(Z)$.

```
6 n {a(X,Z) : b(X)} m  $\leftarrow$  c(Z).
```

Ainsi, dans l'exemple de coloration de nœuds (lignes 7 à 9) : 3 nœuds, a, b et c, sont définis (Ligne 7). 2 signes sont définis de la même manière (ligne 8). Le prédicat de la ligne 9 permettra de générer toutes les combinaisons de colorations pour chaque nœud, soit 27 *answer sets*, un nœud pouvant être coloré "plus", "moins", ou ne pas être coloré.

```
7 noeud(a;b;c).
8 signe(plus;moins).
9 0{coloration(I,S):signe(S)}1 ← noeud(I).
```

2.1.2.3 Agrégations

L'ASP propose aussi des opérations appelées agrégats. Celles-ci s'appliqueront sur un ensemble de prédicats et évalueront une seule valeur à partir d'un poids associé à chacun de ces prédicats. Il en existe 4 types "simples" qui effectueront des opérations sur des valeurs numériques : #sum (somme des poids), #min (poids minimal), #max (poids maximal) et #avg (poids moyen).

Les autres agrégats vont pouvoir exécuter des opérations plus complexes sur le nombre de prédicats, tel que #even (respectivement #odd) qui sera vrai si le nombre de prédicats est pair (respectivement impair). Enfin, l'agrégat utilisé régulièrement dans les méthodes qui seront présentées ensuite : #count qui permettra d'évaluer le nombre de prédicats. Dans l'exemple de la ligne 10 à 14, 27 *answer sets* seront générés, chacun comprenant un ensemble de x prédicats de type "coloration" ($0 \leq x \leq 3$). Le prédicat "totalNodes" (Ligne 13) permettra de les compter (Exemple d'exécution figure 2.1).

```
10 noeud(a;b;c).
11 signe(plus;moins).
12 0{coloration(I,S):signe(S)}1 ← noeud(I).
13 totalNodes(X) ← X=#count{ node(Z) :coloration(Z,L) }.
14 #show totalNodes/1.
```

```
clingo version 4.5.4
Reading from test.lp
Solving...
Answer: 1
noeud(a) noeud(b) noeud(c) signe(plus) signe(moins)
Answer: 2
noeud(a) noeud(b) noeud(c) signe(plus) signe(moins) coloration(a,moins)
Answer: 3
noeud(a) noeud(b) noeud(c) signe(plus) signe(moins) coloration(b,plus)
Answer: 4
noeud(a) noeud(b) noeud(c) signe(plus) signe(moins) coloration(a,moins) coloration(b,plus)
SATISFIABLE

Models      : 4+
Calls       : 1
Time        : 0.001s (Solving: 0.00s 1st Model: 0.00s Unsat: 0.00s)
CPU Time    : 0.000s
```

FIGURE 2.1 – 6 premiers *answer sets* obtenus par le code des lignes 10 à 14.

2.1.2.4 Contraintes et optimisations

Il est possible de limiter les *answer sets* à un sous-ensemble répondant à des critères définis. Ainsi, on peut ajouter une contrainte sur ces *answer sets* via un prédicat sans tête. De la même

manière, il est possible de chercher un ensemble d'*answer sets* qui maximisent ou minimisent un critère. Pour cela, deux opérateurs utilisables : `#maximize` et `#minimize`. Ces opérateurs ont la particularité de permettre la comparaison d'*answer sets* entre eux.

Dans l'exemple des lignes 15 à 21, une contrainte est définie dans le programme logique (ligne 20), qui permettra de ne travailler que sur les solutions contenant le prédicat *coloration(a,plus)*. Autrement dit, les colorations où le nœud *a* sera coloré à *plus*. Une optimisation est définie ensuite (Ligne 21), ce qui permettra de chercher les *answer sets* maximisant la valeur du prédicat "totalNodes", autrement dit ceux ayant un maximum de nœuds colorés.

```

15 noeud(a;b;c) .
16 signe(plus;moins) .
17 0{coloration(I,S):signe(S)}1 ← noeud(I) .
18 totalNodes(X) ← X =#count{ node(Z) :coloration(Z,L) }.
19 #show totalNodes/1.
20 ← not coloration(a,plus) .
21 #maximize {X : totalNodes(X)} .

```

Cette présentation de l'ASP est loin d'être exhaustive, ce langage permettant énormément de possibilités. De plus, de nombreuses nuances existent selon les solveurs utilisés. Néanmoins, les opérateurs montrés précédemment sont ceux qui ont été utilisés dans les méthodes présentées dans les chapitres suivants. De plus, sa syntaxe est assez simple à utiliser pour tester un modèle et le rendre compréhensible. Enfin, les solveurs ASP sont suffisamment efficaces pour utiliser ce langage sur des problèmes complexes, comme des analyses de réseaux biologiques. C'est pour ces raisons que nous avons utilisé ce langage dans les travaux présentés ici.

2.2 Analyse de l'expression des gènes

2.2.1 Introduction

Avec le développement des méthodes de séquençage classique, puis celles à haut débit, la quantité de données biologiques a explosé [88], rendant possible la comparaison de profils d'expression de gènes entre plusieurs individus ou populations. De nombreuses méthodes ont été développées afin de chercher à identifier les gènes d'intérêt à partir de profils d'expression. Nous proposons ici de considérer 2 types d'approches : les approches d'identification de gènes et les approches de discrétisation de gènes.

2.2.2 Identification de gènes

Les approches d'identification de gènes considèrent un ensemble de profils d'expressions appartenant à deux groupes (tissu sain *vs* malade, bon *vs* mauvais pronostic) et chercher à identifier les gènes différentiellement exprimés entre ces 2 groupes. À partir de cet ensemble de profils d'expression de gènes, ces approches renvoient une liste complète de ces gènes, chacun associé à un score d'expression différentielle. C'est sur ce score que peut être appliqué un seuil limite pour considérer l'expression d'un gène comme suffisamment différente entre les 2 groupes d'observations. Bien qu'il existe de très nombreuses manières d'identifier des gènes différentiellement exprimés, quelques-unes sont couramment utilisées afin de comparer 2 distributions

d'expression d'un même gène. L'une des plus classiques est le test-t ou test de student [52] qui permet d'établir la probabilité (p-valeur) que les 2 distributions ne soient pas différentes. Ce modèle se cumule généralement avec une méthode de correction, le plus souvent la correction de Bonferroni [137]. Plusieurs méthodes ont proposé d'améliorer ce modèle statistique, en introduisant une constante basée sur l'écart-type de tous les gènes [132],s permettant ainsi de pouvoir identifier les gènes ayant un niveau d'expression très faible et par conséquent, des variations d'expression réduites. D'autres méthodes se basent sur une comparaison directe des moyennes [33] (*Fold change*) ou des médianes [53](*Median fold change*) en calculant le score *FC* tel que :

$$FC = \frac{\text{Max}(\bar{x}, \bar{y})}{\text{Min}(\bar{x}, \bar{y})}$$

où \bar{x} et \bar{y} représentent les moyennes ou les médianes dans les 2 groupes.

Ces approches considéreront qu'un gène sera différentiellement exprimé si le ratio entre les moyennes (ou les médianes) des 2 populations est supérieur à un seuil. Ce seuil est généralement de 2, on parlera de *two-fold* [91], mais d'autres études utilisent des seuils différents selon l'organisme, le type tissulaire, les groupes, et la méthode de validation associée. Enfin, des approches plus récentes se basent sur des modèles bayésiens [124, 91] permettant d'inférer les gènes différentiellement exprimés en prenant en compte la variance de tous les gènes.

2.2.3 Discrétisation des gènes

Les approches de discrétisation de gènes visent, non plus à identifier les gènes variants entre 2 populations, mais à analyser l'expression des gènes pour chaque profil afin de les associer à une classe (ou cluster). La discrétisation (Figure 2.2) peut être vue comme une fonction f telle que : $f(X) = X'$

où X est une variable continue (qui suppose un nombre infini de valeurs) et X' une variable discrète (qui suppose un nombre fini de valeurs, le plus souvent ordonnées).

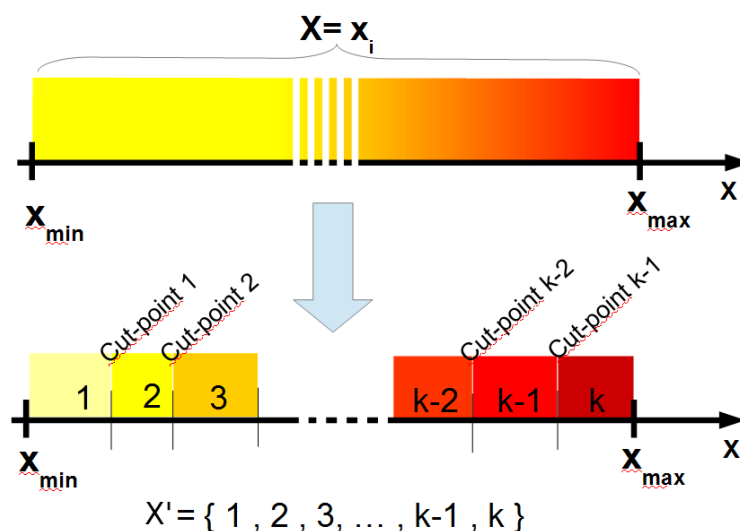


FIGURE 2.2 – Principe de la discrétisation. Image tirée de [35]

Ces approches prennent en entrée un ensemble de n profils d'expressions et retournent n jeux

de données discrétisées. La plupart de ces méthodes sont proches des méthodes de classification non-supervisées, en raison du fait qu'il n'y a que très rarement des informations d'appartenance à une classe/cluster pour les gènes. Il est à noter que le nombre de classes est à déterminer au préalable, cette valeur dépendant de ce qui est associé à chaque cluster (non-exprimé, exprimé, sur/sous-exprimé, fortement sur/sous-exprimé, etc.) Enfin, la discrétisation est une étape primordiale pour énormément d'approches de modélisation [1]. Il existe de très nombreuses approches permettant de discrétiser un jeu de données. Nous en présenterons 5 différentes. Supposons une matrice E de taille $n * m$, n étant le nombre de gènes et m le nombre de profils. X_{ig} représentant l'expression d'un gène g pour le profil i .

2.2.3.1 Equal Width Discretization : EWD

La première méthode, la discrétisation par largeur égale, va viser à créer un ensemble de k clusters d'intervalles égaux pour chaque gène (Figure 2.3-1). Ainsi, l'intervalle de chaque cluster sera d'une taille $\frac{Max(X_g) - Min(X_g)}{k}$. Cette approche, bien qu'intuitive et facile à implémenter, présente de nombreuses limites. Elle reste extrêmement sensible aux valeurs extrêmes (*outliers*), ce qui limite la capacité à discriminer efficacement les niveaux d'expression des gènes [98]. Des méthodes ont été proposées pour réduire cette limitation dans d'autres domaines, en particulier la Winsorisation consistant à définir des quantiles limites et ramener à ces quantiles les valeurs qui sont situées au delà.

2.2.3.2 Equal Frequency Discretization : EFD

La seconde méthode présentée est la discrétisation par fréquence égale, aussi nommée discrétisation par maximisation de l'entropie (*Maximum Entropy Discretization*). Celle-ci vise non plus une égalité d'intervalles entre les clusters mais une égalité en nombre d'observations (Figure 2.3-2). Ainsi, le nombre de profils pour chaque cluster sera de $\frac{m}{k}$. De la même manière que pour la méthode EWD, celle-ci est assez simple. Bien qu'elle soit moins sensible aux données extrêmes, elle présente pour principale limitation le fait que la même valeur observée plusieurs fois peut appartenir à 2 classes [22].

2.2.3.3 Kmeans

La troisième méthode est le partitionnement en k -moyennes (*kmeans*) [21]. Cette approche est incrémentale et va chercher à maximiser l'homogénéité à l'intérieur de chaque classe, et donc l'hétérogénéité entre celles-ci (Figure 2.3-3). Les résultats de cet algorithme dépendront fortement de la méthode de calcul de la distance inter-classe (moyenne des distances, minimum, maximum, médiane, etc.), ainsi que des centres initiaux. Elle présente néanmoins l'avantage de pouvoir être insensibilisée aux valeurs extrêmes (en utilisant des distances médianes) et de générer des classes "cohérentes".

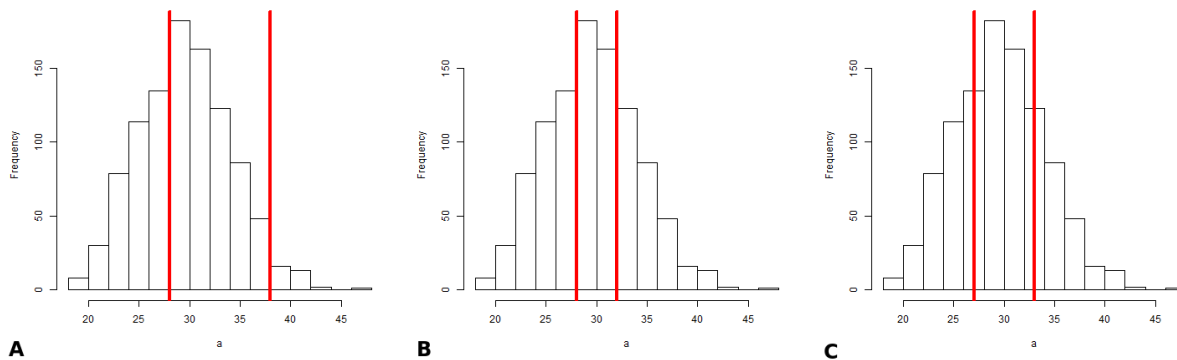


FIGURE 2.3 – Exemple de résultats de discrétisation de données en 3 classes avec deux méthodes différentes. **A** : discrétisation par largeur égale. **B** : discrétisation par fréquence égale. **C** : discrétisation par les K-moyennes

2.2.3.4 *BiKmeans*

La quatrième méthode appelée *BiKmeans* [77] propose une discrétisation plus globale. En effet, celle-ci ne prend pas uniquement en compte la distribution de l'expression des gènes, mais aussi celle des profils d'expression. Elle consiste pour une discrétisation en k classes à réaliser $m * n$ partitionnements en k -moyennes en $k + 1$ classes (numérotées de 1 à $k + 1$). Un partitionnement pour chaque gène et pour chaque profil. La valeur X_{ig} sera discrétisée à la classe x , si le produit des 2 partitionnements est supérieur ou égal à x^2 et inférieur à $(x + 1)^2$. Cette approche, permettant de prendre en compte la variation au sein d'un même profil présente néanmoins plusieurs inconvénients. La répartition issue de cette discrétisation ne sera pas homogène et ceci indépendamment des données. Ainsi, la probabilité d'être dans une petite classe sera bien plus faible que celle d'être dans une classe importante.

Les dernières méthodes que nous présenterons sont les approches par seuil [82]. Elles se rapprochent énormément des approches de *Fold change* utilisées pour l'identification de gènes. Elles visent généralement à identifier les gènes sur/sous-exprimés dans chaque profil, en fixant un seuil (la moyenne ou la médiane de l'expression du gène). Si un profil i présente un niveau d'expression g supérieur (respectivement inférieur) à ce seuil, on pourra alors considérer X_{ig} comme sur-exprimé (respectivement sous-exprimé). D'autres approches proposent de discrétiser en 3 états possibles [36] en ajoutant la non-variation comme information. Ces seuils pourront se baser sur la dispersion de l'expression du gène (généralement l'écart-type), ou être fixés préalablement. Enfin, d'autres approches vont discrétiser en un nombre plus important de classes, en 5 [129], afin d'intégrer des variations "ambiguës", pas assez fortes pour être sous/sous-exprimées mais pas assez faibles pour être considérées comme invariantes.

2.3 Myélome multiple

2.3.1 Différenciation normale du plasmocyte

Les anticorps sont l'un des piliers de l'immunité acquise (adaptative, évolutive et spécifique) permettant à l'organisme de se défendre face à un nouvel agent pathogène. Les anticorps sont

produits par les plasmocytes, cellules dérivées des lymphocytes B, eux-mêmes issus des cellules souches lymphoïdes. Ces cellules souches, par le processus de lymphopoïèse, vont se différencier en cellules pro-B. C'est à ce stade que des séries de réarrangements des gènes codants pour les régions variables des chaînes lourdes puis des chaînes légères vont permettre l'expression en surface d'une immunoglobuline unique.

Ces cellules B désormais matures vont migrer au niveau des organes lymphoïdes secondaires, où elles pourront fixer un antigène compatible avec l'immunoglobuline de surface. Cette fixation activera des voies de signalisation diverses (apoptose, prolifération). Ces activations déclencheront une nouvelle série de recombinaisons génétiques, afin d'améliorer la spécificité antigénique. Enfin, ces lymphocytes B se différencieront en 2 types cellulaires : les cellules B mémoires, qui passeront dans le sang et circuleront dans tout l'organisme, et les plasmocytes qui migreront principalement dans la moelle osseuse et dont une partie deviendra des plasmocytes à longue durée de vie grâce aux signaux de l'environnement.

La différenciation des plasmocytes est donc inféodée à de nombreuses voies de signalisation, amenant des recombinaisons génétiques lourdes. Un dysfonctionnement de ces processus peut être en partie à l'origine de l'apparition de plasmocytes tumoraux en touchant des gènes régulateurs de la prolifération tels que *CCND1* [122] ou *RB1* [20].

2.3.2 Aspect clinique et épidémiologie

Le myélome multiple (ou syndrome de Kahler), touchant les plasmocytes, se caractérise par la multiplication dans la moelle osseuse de ceux-ci et semble avoir été décrit pour la première fois en 1840 sur un patient anglais [73].

L'évolution de cette maladie passe par plusieurs phases notamment une gammopathie monoclonale de signification indéterminée (MGUS) et un myélome asymptomatique [95]. Ces étapes sont détectées de façon fortuite car asymptomatiques. La probabilité d'évolution d'un myélome asymptomatique (20% des myélomes totaux) à une phase symptomatique est de 10% par an [81]. Enfin, un myélome multiple, en phase finale, peut évoluer en leucémie par le passage des cellules tumorales dans le sang.

Le myélome multiple est diagnostiqué par une électrophorèse des protéines sériques et confirmé principalement par un myélogramme. Ces examens visent à identifier la sur-production d'une immunoglobuline monoclonale (produite par le plasmocyte tumoral) et l'envahissement tumoral.

Les symptômes cliniques sont divers, du fait de l'hétérogénéité de cette pathologie. Néanmoins, certains de ces symptômes, lourdement handicapants, sont utilisés pour le diagnostic : Les plasmocytes cancéreux peuvent déréguler le cycle de remodelage osseux (la balance ostéoblaste/ostéoclaste), entraînant une dégradation osseuse accélérée. Cette dégradation peut provoquer une ostéoporose, des lésions ostéolytiques, visibles à l'imagerie radio. Ce dérèglement est également à l'origine de douleurs osseuses, de tassements vertébraux, et de fractures osseuses. Cette dégradation peut aussi être à l'origine d'une hypercalcémie, elle même provoquant une viscosité anormalement forte du sang. De la même manière, la production des cellules sanguines peut être perturbée, déclenchant anémie, neutropénie et thrombopénie. La surproduction de l'immunoglobuline par les cellules cancéreuses peut être à l'origine d'une déficience des

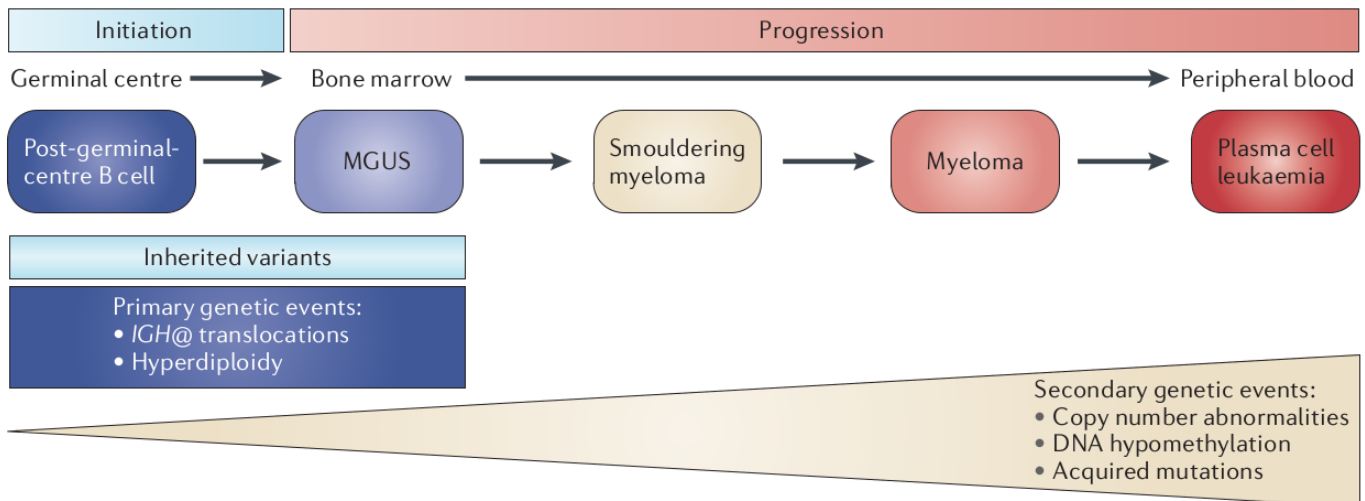


FIGURE 2.4 – Initiation et progression du myélome multiple (tirée de [95]).

fonctions rénales par son dépôt dans les reins. Enfin, la prolifération du plasmocyte cancéreux peut empêcher le développement des autres plasmocytes affaiblissant ainsi le système immunitaire.

En 2012, le myélome multiple avait un taux annuel d'incidence standardisé de 8.3 pour les hommes et de 5.3 pour les femmes en France contre respectivement 4.7 et 3.1 en Europe [41]. Il représente un peu plus de 1% des cancers, et 2% de la mortalité associée [113] avec un taux de survie d'environ 49.6% après 5 ans. Il existe des prédispositions inter et intra-population. D'un point de vue des populations : les hommes semblent plus souvent touchés que les femmes. De la même manière, une étude américaine a évalué l'incidence sur la population afro-américaine comme 2 fois supérieure à celle euro-américaine [135]. Enfin l'âge médian de diagnostic du myélome multiple est de 70 ans en France, et de 65 aux États-Unis d'Amérique.

2.3.3 Évolution et facteurs pronostiques

Il existe de nombreux critères identifiés comme marqueurs de l'avancement et de la gravité de la pathologie chez un patient. Le niveau d'avancement de la tumeur (MGUS, myélome multiple asymptomatique et myélome multiple) est évalué, entre autres, par la concentration de la protéine monoclonale, l'apparition de symptômes cliniques et l'envahissement tumoral dans la moelle. D'autres marqueurs ont pu être identifiés mais ne sont pas utilisés en raison du caractère invasif et coûteux des analyses. Ils ont néanmoins permis de mieux comprendre les mécanismes impliqués dans cette maladie.

Au cours de son développement, de nombreuses modifications vont se cumuler dans les cellules cancéreuses. En effet, bien que l'on considère le myélome multiple comme une seule maladie, la diversité de ces mutations entre les cellules cancéreuses rend cette pathologie bien moins homogène, et ceci même au sein de la tumeur [83, 87]. Ces modifications ont des impacts sur les mécanismes intra et inter-cellulaires, permettant le maintien du développement des cellules cancéreuses. Elles peuvent se présenter à diverses échelles : chromosomiques (chromo-

some(s) surnuméraire(s), gains et délétion de bras chromosomiques, translocations) [26, 95], épigénétiques et génétiques. Celles-ci peuvent être étudiées par des analyses cytogénétiques et de séquençage. Les mutations touchant un seul gène sont très fréquentes bien que très variées. Ces mutations, non-sens dans la majorité des cas [18], entraînent le plus souvent une diminution de l'expression du gène associé [114], suggérant que la plupart de ces mutations ont un effet très réduit sur les mécanismes cellulaires. Néanmoins, il est à noter la redondance de mutations de gènes impliqués dans le cycle cellulaire, la survie, ou la résistance à l'apoptose [95]. L'analyse de ces anomalies dans les tumeurs a permis d'identifier des voies impliquées dans le myélome multiple. C'est le cas, par exemple, des voies Ras/PAPK et NFκB dont respectivement 43% et 17% des patients sont porteurs de mutations ponctuelles [134, 4]. Des marqueurs protéiques ont aussi été identifiés tels que la protéine RB1 [103] ou le facteur de transcription FOXM1 [133, 50]. Notons par ailleurs que ces marqueurs ne sont pas toujours associables à des anomalies chromosomiques ou des mutations.

Certaines de ces modifications, redondantes, ont pu être identifiées comme facteurs de bon ou mauvais pronostic pour les patients atteints. Sur le plan des modifications chromosomiques, si certaines, comme la trisomie 3, semblent indiquer un meilleur pronostic, d'autres, comme la trisomie 21, sont de mauvais signe [26]. Deux autres modifications chromosomiques, très connues, sont la délétion du bras court du chromosome 17 (del(17p)), et la translocation entre les chromosomes 4 et 14 (t(4;14)) présentes dans, respectivement, 11% et 14% des cas [5]. Ces mutations sont associées à un facteur de risque supérieur. A l'échelle génétique, les gènes mutés peuvent aussi impacter la survie des patients. Les patients à faible pronostic semblent avoir des gènes sur-exprimés impliqués dans le cycle cellulaire [32].

Le myélome multiple, de part son hétérogénéité, est une pathologie extrêmement difficile à évaluer et donc à combattre. Bien que les analyses cytogénétiques et géniques permettent d'avoir une première évaluation de la gravité et de l'intérêt potentiel de l'analyse des gènes codants, des longs ARN non codants, et des séquences régulatrices, leur impact au niveau de l'expression du gène semble être faible. Ces analyses restent ainsi limitées dans l'explication des mécanismes sous-jacents dans les processus cellulaires impliqués. Les approches d'analyses du transcriptome présentent l'avantage d'oblitérer ces mutations "spectatrices" [114], en étudiant directement l'impact de celles-ci sur l'expression des gènes des cellules tumorales. De plus, l'analyse du transcriptome chez un patient est une procédure bien moins invasive que pour celle du protéome. Ces analyses de l'expression des gènes, si elles ont initialement visé à identifier des gènes comme facteurs pronostiques, se sont enrichies des connaissances biologiques en intégrant les fonctions et/ou les interactions associées à ces gènes permettant d'identifier des voies de signalisation perturbées [66]. Ainsi, il est aujourd'hui possible d'ajouter une dimension fonctionnelle et donc potentiellement causale à l'identification de ces perturbations, permettant ainsi d'aller vers des approches de recherches thérapeutiques.

2.4 Des gènes à la fonction : les *Pathway Analysis*

2.4.1 Introduction

Avec le développement des technologies de séquençages, et parallèlement l'augmentation des connaissances biologiques, beaucoup de modèles cherchant à intégrer des données biologiques (expression de gènes, phospho-protéomique, etc.) avec des réseaux ([67]) se sont développées. Ces méthodes se basent sur des agglomérations de gènes ou protéines sur la base de fonctions biologiques communes afin d'identifier les processus biologiques mis en œuvre ou dérégulés dans les populations étudiées. Ces méthodes sont regroupées au sein des *Pathway analysis* et visent à caractériser 2 groupes/classes de données (malades *vs* sains, bons *vs* mauvais pronostic, etc.) en confrontant des données issues de ces 2 groupes avec des informations provenant d'une base de données de régulation. Si la plupart cherche à identifier des éléments-clés pouvant expliquer les variations entre ces classes, les approches utilisées peuvent diverger fortement. Nous pouvons néanmoins en considérer 3 grands types selon Khatri *et al.* [67] : les ORA (Over-representation analysis), les FCS (Functional Class Scoring) et les PT (pathway topology).

TABLE 2.1 – Méthodes d'intégration base de données/Observations ORA et FCS. *ORA* : over-Representation Analysis, *FCS* : functional Class Scoring. *FA* : forward assumption (logique de conséquence), *BA* : backward assumption (logique de causalité).

Nom de la méthode	Type	Logique	Type de données	Base de données
Panther [92]	<i>ORA</i>	<i>FA</i>	Qualitative	GO
Cluego [15]	<i>ORA</i>	<i>FA</i>	Qualitative	KEGG
GSEA [126]	<i>FCS</i>	<i>FA</i>	Quantitative	MSigDB
MARINa [75]	<i>FCS</i>	<i>BA</i>	Quantitative	HBCI

2.4.2 Over-Representation Analysis (ORA)

Ce groupe contient les approches basées sur les gènes différentiellement exprimés (DE). La plupart des ORA (Tableau 2.1), en utilisant des bases de données tel que Gene Ontology [127] ou KEGG, considèrent les protéines issues de ces gènes DE pour faire l'association entre gènes et des entités appelées *annotations*. Ces *annotations* peuvent caractériser une fonction biologique (exemple : apoptose), une fonction moléculaire (exemple : fixation sur le récepteur CD40) ou la localisation (exemple : membrane plasmique). Enfin, les ontologies sont connectées entre elles par des relations hiérarchisées (Figure 2.5). Dans notre cas, nous ne travaillerons que sur les fonctions biologiques, que l'on peut considérer comme des *pathways*.

Dans les méthodes ORA, chaque *annotation* va se voir attribuer un score en fonction de la proportion de gènes DE contenus dans celle-ci. Les gènes DE peuvent être identifiés avec un test statistique ou avec un seuil. De même, l'attribution du score peut se faire par plusieurs méthodes, la plus utilisée est le test hypergéométrique [15] qui compare le nombre de gènes dans une *annotation* par rapport à une distribution aléatoire. Les approches d'ORA diffèrent

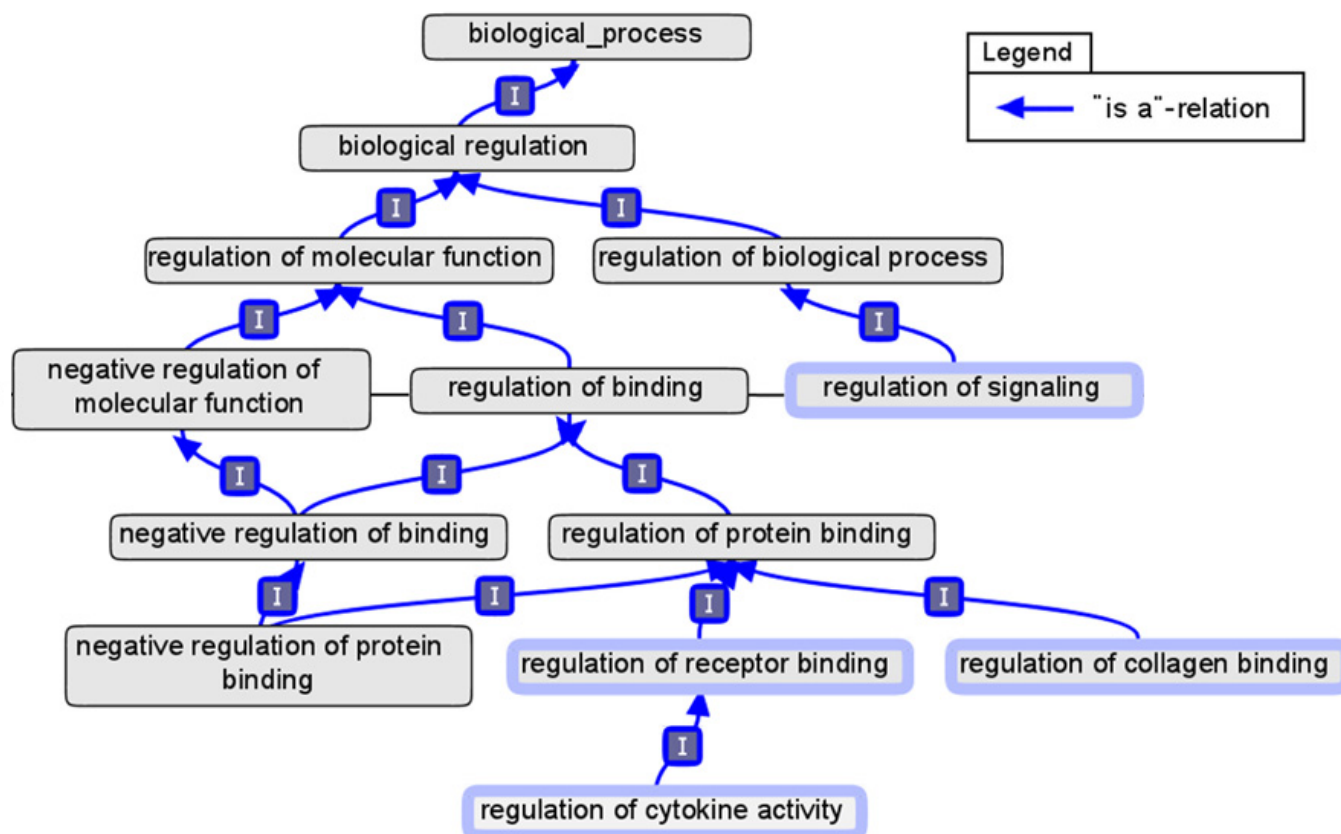


FIGURE 2.5 – Exemple d'ontologies et de leurs relations. Image issue de [39] avec modifications.

entre elles par la combinaison de la méthode d'identification des DE, l'évaluation des *pathways* ainsi que la base de données contenant l'information sur ces *pathways*.

Martin *et al.* [86] considère ce type de raisonnement comme fonctionnant avec une logique de conséquence (*forward assumption*). Cette logique se base sur la production d'un gène pour l'associer à un ou plusieurs *pathways* par comparaison avec des approches basées sur une logique de causalité (*backward assumption*) considérant les *pathways* comme cause des gènes DE. Néanmoins, ces méthodes présentent de nombreuses limitations, dues au fait que l'identification des gènes DE réduit énormément l'information présente. De plus, les *pathways* sont étudiés indépendamment les uns des autres, malgré leurs interconnexions connues. Enfin, l'estimation de la différence d'expression des gènes ne prend pas en compte le sens de cette différence. Ainsi un gène sur-exprimé sera considéré comme ayant le même impact que s'il était sous-exprimé, malgré un effet très potentiellement différent.

Dans cette thèse (chapitre 4), nous utiliserons l'implémentation d'une méthode de type ORA, appelée PANTHER [92]. Cet outil, disponible en ligne à partir de la base Gene Ontology (GO, <http://geneontology.org/>) prend en entrée une liste de gènes DE, et identifie les *pathways* (on parlera aussi de termes GO) sur et sous-représentés par rapport à une distribution aléatoire (Figure 2.6). Le test utilisé est un test binomial (comparaison de fréquences) associé à une correction de Bonferroni. Cette correction permet de prendre en compte la multitude de

tests statistiques effectués (un test pour chaque *pathway* étudié) et de limiter le nombre de faux-positifs. Pour chaque *pathway* i , le pourcentage (P_i) de gènes de la base de données associés à celui-ci va être calculé. Il est possible alors, à partir d'une liste de n gènes (fournie par l'utilisateur) et de ce pourcentage, de calculer le nombre théorique (T_i , Figure 2.6, colonne "expected") de gènes associés à ce *pathway*. On aura $T_i = P_i * n$. Il est alors possible de comparer cette valeur théorique avec celle observée dans la liste (O_i , Figure 2.6, colonne "#") et de calculer la probabilité (Figure 2.6, colonne "P-value") que les gènes associés à ces *pathways* de la liste entrée par l'utilisateur soient sur ou sous-représentés (Figure 2.6, colonne "+/-") par rapport à une proportion théorique.

GO biological process complete	Homo sapiens (REF)				
	#	#	expected	Fold Enrichment +/-	P value
ovulation from ovarian follicle	9	4	.10	41.99	+ 2.65E-02
↳ gonad development	207	15	2.19	6.85	+ 7.64E-05
↳ single-organism developmental process	5380	145	56.95	2.55	+ 2.59E-31
↳ developmental process	5467	146	57.87	2.52	+ 3.21E-31
↳ single-organism process	12702	212	134.46	1.58	+ 3.72E-30
↳ reproductive structure development	416	29	4.40	6.59	+ 1.61E-11
↳ anatomical structure development	5104	138	54.03	2.55	+ 7.01E-29
↳ reproductive system development	420	29	4.45	6.52	+ 2.05E-11
↳ system development	4174	124	44.18	2.81	+ 3.49E-28
↳ multicellular organism development	4760	133	50.39	2.64	+ 1.20E-28
↳ single-multicellular organism process	5554	145	58.79	2.47	+ 1.11E-29
↳ multicellular organismal process	6635	151	70.24	2.15	+ 1.06E-24
↳ developmental process involved in reproduction	627	35	6.64	5.27	+ 1.17E-11
↳ reproductive process	1351	46	14.30	3.22	+ 1.58E-08
↳ reproduction	1351	46	14.30	3.22	+ 1.58E-08
↳ development of primary sexual characteristics	212	15	2.24	6.68	+ 1.04E-04
↳ sex differentiation	260	16	2.75	5.81	+ 2.29E-04
↳ single organism reproductive process	1211	44	12.82	3.43	+ 6.16E-09
↳ animal organ development	2966	100	31.40	3.19	+ 1.22E-24
↳ ovulation	20	5	.21	23.62	+ 2.44E-02
↳ multicellular organismal reproductive process	793	28	8.39	3.34	+ 2.69E-04
↳ multicellular organism reproduction	803	29	8.50	3.41	+ 9.17E-05
↳ multi-organism reproductive process	943	30	9.98	3.01	+ 8.02E-04
↳ multi-organism process	2297	85	24.31	3.50	+ 2.41E-22
↳ rhythmic process	288	17	3.05	5.58	+ 1.52E-04
↳ ovulation cycle	102	9	1.08	8.34	+ 1.57E-02

FIGURE 2.6 – Exemple d'analyse avec l'outil PANTHER sur une liste de 222 gènes. Les *annotations* sont triées par ordre hiérarchique.

2.4.3 Functional Class Scoring (FCS)

Ce type de méthodes utilise l'intégralité des données sans pré-sélection permettant d'intégrer l'effet de petites variations d'expression de gènes pour identifier les *pathways* impliqués. Il existe plusieurs méthodes appartenant aux FCS (Tableau 2.1) : Les *Gene Set Enrichment Analysis* (GSEA) [126], *GeneTrail* [6] ou *the MAster Regulator INference algorithm* (MARINA) [75]. Ces méthodes sont menées en 3 étapes. Premièrement, chaque gène est évalué selon le niveau de différence d'expression entre les classes que l'on veut caractériser. Ensuite, chaque *pathway* est évalué selon le score des gènes présent dans celui-ci. Enfin, la significativité de chaque

pathway (*significance level* : *SL*) peut être estimée. Les approches FCS peuvent avoir un raisonnement basé sur une logique de conséquence [126, 6] ou de causalité [75, 71]. Si ces méthodes améliorent le problème lié à la sélection des gènes, les *pathways* sont toujours étudiés indépendamment entre eux. De plus, nous pouvons noter que la position des gènes dans un *pathway* ainsi que leurs interactions ne sont pas prises en compte. Tout comme les ORA, les différentiels d'expression ne prennent pas non plus en compte le type de différences (sur/sous-exprimé), ce qui limite l'analyse de l'impact sur les *pathways*.

2.4.4 Pathway Topology (PT)

Ces méthodes prennent en compte les interdépendances entre les gènes à partir de la topologie des *pathways* (Tableau 2.2). Certaines de ces approches vont se baser sur des réseaux d'interactions non orientés, en particulier les réseaux *Protein-protein interactions* (PPIs) [115, 39] tandis que d'autres vont intégrer une notion d'orientation dans les interactions : on parlera alors de réseau de régulation. Quelques approches PT intègrent aussi les différents types d'interactions possibles entre ces gènes [86], généralement l'activation et l'inhibition. Néanmoins, ces approches restent minoritaires. De même que pour les méthodes ORA et FCS, nous trouvons des logiques de conséquence [59, 70, 79, 143, 39, 7] et de causalité [34, 86]. Quelques méthodes PT continuent à analyser indépendamment chaque *pathway* en leur associant un score selon les gènes de ces *pathways* et leur position dans ceux-ci [34, 86]. D'autres intègrent tous les *pathways* en même temps et prennent ainsi en compte les connexions entre eux. L'évaluation de ces *pathways* peut alors se faire, soit par des approches de marches aléatoires, en prenant en compte l'orientation des interactions [79], permettant ainsi d'identifier les gènes ayant le plus d'impact sur les autres gènes, tandis que d'autres considéreront les interactions comme bidirectionnelles [70], afin d'identifier des communautés de gènes. D'autres approches sont associées à la théorie des jeux [115]. Certaines méthodes de PT, elles, chercheront à identifier un sous-graphe par optimisation dans des réseaux PPIs [39] ou de régulation [7]. Nous pouvons aussi signaler des approches PT utilisant l'information des profils d'expression ainsi que celle des réseaux de régulation pour inférer l'état des facteurs de transcription régulant les gènes analysés. Il est alors possible, non plus d'étudier l'expression des gènes, mais celle de leurs régulateurs [86, 129, 36]. Enfin, la méthode que nous avons utilisée dans cette thèse se base sur la coloration cohérente des graphes. Dans la section suivante, nous présenterons en détail ce modèle et son implémentation *Iggy* [129].

2.5 Méthode de coloration cohérente des graphes

2.5.1 Introduction

Les modèles de coloration de graphes visent à confronter un ensemble d'observations (issues de données transcriptomiques ou protéomiques) avec un graphe de régulation afin d'identifier les incohérences entre ceux-ci et inférer l'état des éléments non-observés. Il est alors possible de réparer soit le graphe, soit les observations afin de rétablir la cohérence. Le premier modèle proposé, implémenté via l'outil *Ingranalyze* [46], utilise une modélisation à 2 signes, c'est-à-

TABLE 2.2 – Vue globale des méthodes de type *Pathway topology* présentées. *FA* : forward assumption (logique de conséquence), *BA* : backward assumption (logique de causalité).

Nom de la méthode (ou de l'auteur.e)	Logique	<i>Pathways</i> analysés individuellement	Interactions orientées	Activations et inhibitions	Type de données	Base de données
DRW [79]	FA	✗	✓	✗	Quantitative	KEGG
NetWalker [70]	FA	✗	✗	✗	Qualitative	multi-bdd
Impact-analysis [34]	BA	✓	✓	✗	Quantitative	KEGG
TopoNPA [86]	BA	✓	✓	✓	Qualitative	CBN
Razi [115]	FA	✗	✗	✗	Quantitative	Human PPI network
Faisal [39]	FA	✗	✗	✗	Qualitative	multi-bdd
Backes [7]	FA	✗	✓	✗	Quantitative	KEGG
Licorn [36]	FA	✓	✓	✓	Qualitative	multi-bdd
Iggy [129]	FA/BA	✗	✓	✓	Qualitative	NCI-PID

dire qui représente des observations d'activations (+) et d'inhibitions (-). Le second, utilisé dans cette thèse, implémenté via *Iggy* [129], est un modèle à 3 signes, en ajoutant les observations de type "invariant" (0).

Ces 2 modèles proposent, à partir de cette confrontation graphe-observation, de calculer les ensembles minimaux de réparations (du graphe ou des observations) afin de restaurer la cohérence, puis de déduire l'état des éléments du graphe non-observé (avec réparation si nécessaire). Enfin, *Iggy* ainsi que son prédécesseur (*Ingranalyze*) utilisent une approche d'exploration exhaustive, par programmation logique (ASP : *Answer Set Programming*) de toutes les colorations du graphe, en tenant compte des observations, ainsi que de règles de cohérence, garantissant que l'état d'une entité du graphe peut être expliqué par les prédécesseurs.

La coloration cohérente des graphes à 3 signes a été utilisée, via l'outil *Iggy*, initialement sur des données issues d'*E.Coli* avec un réseau obtenu à partir de RegulonDB [129]. Une seconde application d'*Iggy* a été faite sur lignées humaines de kératinocytes et de fibroblastes en les confrontant à des réseaux issus de la base de données PID [68].

Pour la suite, nous ne travaillerons que sur la méthode de coloration cohérente des graphes intégrant les observations variantes (+ et -) et non-variantes (0).

2.5.2 Coloration d'un graphe

Supposons un graphe orienté $G(V, E, \alpha)$, où V représente l'ensemble des nœuds, E , l'ensemble des arcs et α , une fonction associant chaque arc à un signe tel que $\alpha : E \rightarrow \{+, -\}$. Cette fonction permet de modéliser les 2 types d'interactions utilisées dans ce modèle : activation et inhibition. Dans le cas d'un arc $s \rightarrow n$ partant de s vers n et associé à un signe "+", on considérera que s est un activateur de n . Dans le cas où un arc est associé à un signe "-", on le définira comme un inhibiteur de n .

Le modèle de coloration d'un graphe considère qu'il existe, au moins, une configuration ou *une coloration* de ce graphe. Une coloration est l'association d'un signe $\{+, -, 0\}$ pour tous les nœuds de V . Définissons S comme l'ensemble des colorations du graphe G possibles, on notera

que $|S| = 3^{|V|}$.

2.5.3 Règles de cohérence des signes

On considère un nœud $n \in V$ comme cohérent, si sa coloration peut être "expliquée" par, au moins un des prédécesseurs de n dans G . Cette notion peut être implémentée avec des règles de cohérences des signes pouvant être vérifiées automatiquement :

1. Tous les nœuds considérés comme *input* sont cohérents. Généralement, on considère comme *input* les nœuds sans prédécesseur.
2. Chaque variation $\{+, -\}$ associée à un nœud n dans une coloration de graphe donnée doit être expliquée par un prédécesseur de n . Ainsi, chaque nœud variant (associé à un signe $\{+, -\}$) qui n'est pas un input, doit au moins avoir un activateur du même signe ou un inhibiteur du signe opposé (+ si n est associé à - et inversement).
3. Chaque nœud invariant n (associé à un signe 0) doit être expliqué le fait que soit (i) tous les prédécesseurs de n sont associés à 0, soit (ii) deux prédécesseurs de n ont des influences variantes contraires.

On considérera une coloration du graphe G comme cohérente si tous les nœuds de ce graphe sont cohérents (figure 2.7).

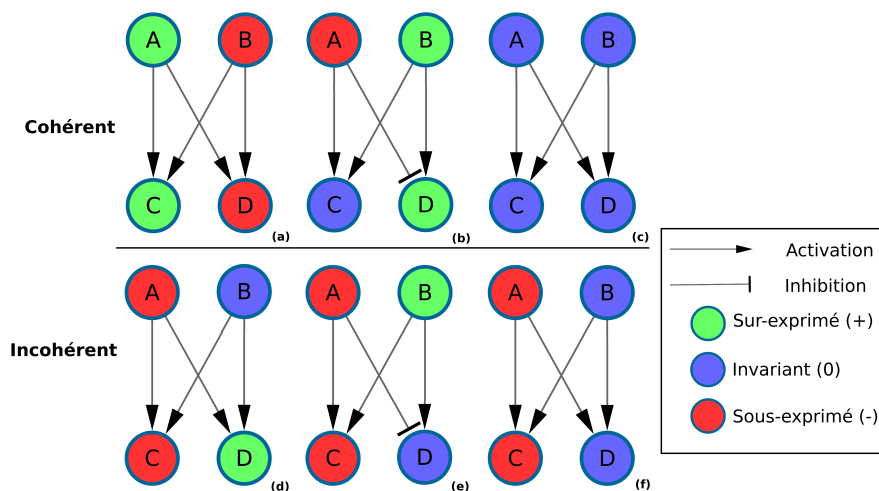


FIGURE 2.7 – Exemple de colorations cohérentes (a,b,c) et incohérentes (d,e,f). Dans les colorations (d),(e), et (f), le nœud D est incohérent. Dans la coloration (d), le nœud D ne respecte pas la règle 2. Les colorations (e) et (f) sont incohérentes car le nœud D ne respecte pas la règle 3.

2.5.4 Observations

Supposons β , comme une fonction associant des nœuds du graphe G à un signe tel que $\beta : V \rightarrow \{+,-,0\}$. Généralement, β peut être considéré comme un ensemble d'observations sur une partie

des nœuds du graphe. Ainsi, il restera des nœuds qui ne seront pas associés à un signe. En considérant β , on peut définir (S^*) le sous-ensemble des colorations de G respectant β . On notera que $|S^*| = 3^{|V| - |\beta|}$

2.5.5 Réparations

Dans le cas où il n'y a pas de coloration cohérente de G contenant β , il est possible de réparer le modèle de 2 manières : soit en corrigeant G (MCOS-repair) ou en corrigeant β (SCENFIT-repair)

MCOS-repair : Ce type de réparation corrige la topologie du graphe. En effet, on peut considérer que G est incomplet, que ce soit dû à des informations manquantes, ou à la méthode de génération. On peut donc supposer que certaines incohérences entre G et β sont causées par des éléments manquants dans le graphe. Il est alors possible de corriger celui-ci en ajoutant des influences artificielles (figure 2.8). Dans ce cas, on cherchera l'ensemble de corrections de taille minimale (*cardinal minimal correction set* : MCOS) afin de restaurer la cohérence entre G et β . Notons qu'il existe, généralement, plusieurs manières de corriger un graphe, aussi le MCOS n'est généralement pas unique.

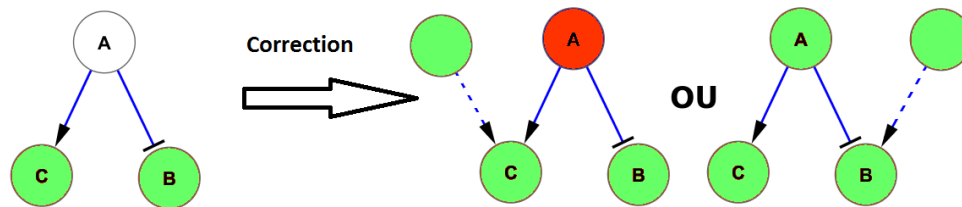


FIGURE 2.8 – Exemple de réparations de type MCOS avec l'ajout d'une influence artificielle pour chacune.

SCENFIT-repair : Ce type de réparation va corriger β en considérant qu'il y a de fausses informations dans les données observées. β sera corrigé en changeant le signe des nœuds observés (figure 2.9). Chaque type de changement se verra attribuer un coût qui devrait être minimal pour restaurer la cohérence.

1. Changer un signe variant (+,-) vers le signe variant opposé aura un coût de 2.
2. Changer un signe invariant (variant) vers un signe variant (invariant) aura un coût de 1.

On peut noter que là aussi, il peut y avoir plusieurs manières de réparer β pour restaurer la cohérence. Afin de réduire ce nombre de possibilités de corrections, le coût total (soit la somme des coûts de chaque correction locale) devra être minimale.

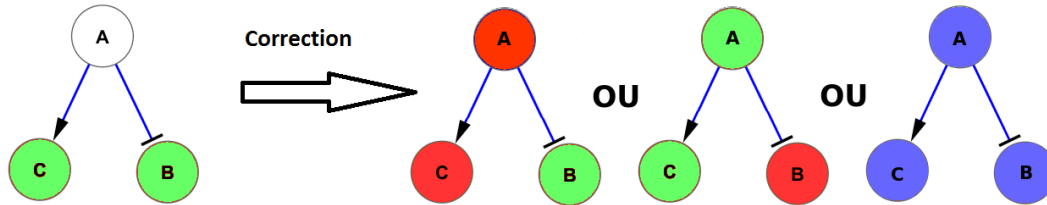


FIGURE 2.9 – Exemple de réparations de type SCENFIT avec un coût total pour chacune de réparation de 2.

2.5.6 Prédictions et projections

Après identification des ensembles de réparations (MCOS ou SCENFIT), l'ensemble des colorations sera l'union des colorations du graphe. Généralement, le nombre de colorations cohérentes du graphe est très important, aussi nous utilisons une projection de celles-ci pour en extraire l'information à partir d'une confrontation graphe-observations. On peut considérer 7 classes de projections possibles, réparties en 3 types :

- 3 classes sont de type *forte*, signifiant que le nœud associé à une prédiction de cette classe a le même signe dans toutes les solutions cohérentes $\{+,-,0\}$. Par exemple, dans la figure 2.8, le nœud C est associé à + dans toutes les solutions.
- 3 classes sont de type *faible*, signifiant que le nœud associé à une prédiction de cette classe est associé à 2 signes dans les solutions cohérentes : *Not+* $(-,0)$, *Not-* $(+,0)$, *change* $(+,-)$. Ainsi, dans la figure 2.8, le nœud A est associé à + et à - dans les solutions cohérentes, il sera donc associé à la classe "change".
- La dernière classe porte sur les nœuds qui seront associés aux 3 signes dans les solutions cohérentes ? $(+,-,0)$. C'est le cas du nœud A dans la figure 2.9.

Modèle de coloration des graphes pour le myélome multiple

3.1 Introduction

L'Un des premiers objectifs de cette thèse était de modéliser, grâce à la coloration cohérente des graphes, des données issues de patients atteints du myélome multiple afin de pouvoir les comparer. En effet, les précédentes applications sur des lignées humaines portaient sur des lignées de cellules saines immortalisées (kératinocyte et fibroblastes) [68] confrontées à des réseaux de signalisation issus de la base PID.

Dans cette contribution, nous avons travaillé avec un ensemble de données d'expression de gènes, issu d'échantillons de plasmocytes provenant de 611 individus, 602 malades (MC : Myeloma cells) et 9 normaux (NPC : Normal plasma cells) fournies par l'équipe 11 du CRCINA. Après extraction des ARN, les niveaux d'expression de gènes ont été mesurés par des puces Affymetrix Human Exon1.0

3.2 Traitement des données

3.2.1 Discrétisation des données

Du fait que notre modèle de coloration de graphes utilise les expressions invariantes de gènes, les méthodes de discrétisation classiques ne pouvaient être utilisées, néanmoins la dernière application sur des lignées humaines [68] proposait une méthode de discrétisation que nous avons reprise. Pour identifier des gènes sur/sous-exprimés et invariants pour chaque profil d'expression de gènes (GEP), nous avons utilisé 2 seuils : k_1 pour les gènes invariants et k_2 pour les gènes variants (Figure 3.1). Nous avons calculé pour chaque gène g , un vecteur p^g , composé des valeurs p_i^g pour chaque GEP i . Les valeurs p_i^g ont été calculées en soustrayant à chaque

GEP i , le niveau d'expression moyenne de g chez les données normales. Ensuite, nous avons discrétisé les valeurs de p_i^g en utilisant 2 seuils k_1 et k_2 :

Si $p_i^g > k_2$, g est considéré comme sur-exprimé pour i ;

Si $p_i^g < -k_2$, g est considéré comme sous-exprimé pour i ;

Et si $-k_1 < p_i^g < k_1$, g est considéré comme invariant pour i .

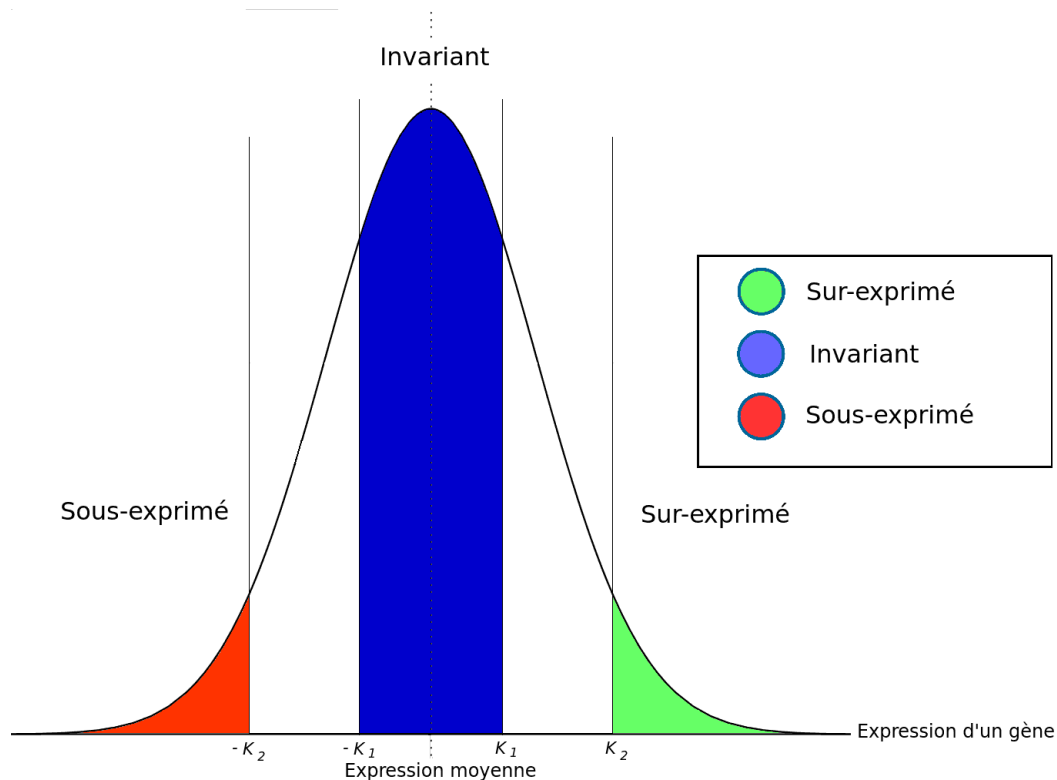


FIGURE 3.1 – Méthode de discrétisation avec les seuils k_1 et k_2 .

En choisissant plusieurs combinaisons de valeurs pour k_1 et k_2 , nous avons obtenu 150 ensembles de données contenant les gènes sur-exprimés (+), sous-exprimés (-) et invariant (0). Les valeurs utilisées pour k_1 et k_2 étaient :

- $k_1 \in \{0, 0.01, 0.02, 0.03, 0.04, 0.05, 0.1, 0.15, 0.2, 0.3\}$ pour les gènes invariants.
- $k_2 \in \{0.2, 0.4, 0.6, 0.8, 1, 1.2, 1.4, 1.6, 1.8, 2, 2.2, 2.4, 2.6, 2.8, 3\}$ pour les gènes variants.

Comme premier filtre, nous avons rejeté les ensembles dont la proportion d'un des signes dépassait 50%. Ensuite, nous avons calculé la précision de chaque combinaison de seuils. Pour chaque ensemble de données, nous avons utilisé 100 fois 50% des données discrétisées $\{+, -, 0\}$ afin de prédire les 50% restants. Le calcul de la précision a ensuite été effectué en utilisant une matrice de précision (Tableau 3.1) afin de prendre en compte les classes de prédictions faibles (*not+*, *not-*, *change*).

La combinaison de seuils permettant la meilleure précision de 43% (IC 95% : $\pm 3\%$) était $k_1 = 0.03$ and $k_2 = 0.2$. Nous avons utilisé l'ensemble associé à cette combinaison pour le reste de l'application. Notons aussi que nous avons aussi tenté une approche de discrétisation par

TABLE 3.1 – Matrice de précision utilisée pour les calculs de précision des prédictions incluant les signes +,- et 0.

Prédictions	Signe observé		
	+	-	0
+	1	0	0
-	0	1	0
0	0	0	1
CHANGE	0.5	0.5	0
Not+	0	0.5	0.5
Not-	0.5	0	0.5
?	0.3	0.3	0.3

K-means ($k=3$), mais la précision restait statistiquement inférieure à la combinaison identifiée. De la même manière, afin de valider l'apport de l'intégration des gènes invariants, nous avons calculé la précision des prédictions uniquement avec les gènes variants (sur et sous-exprimés). Enfin, pour démontrer l'intérêt d'utiliser les gènes invariants, nous avons calculé la précision sur un modèle à 2 signes (sur/sous-exprimé) en utilisant 50% des observations. Ce modèle prend donc en entrée et ne prédit que des colorations de sur et sous-expression. La distribution aléatoire des prédictions donnera une précision de 50% (Table 3.2). La précision calculée était de 48%, valeur très proche et inférieure à celle d'une précision aléatoire tandis que celle obtenue avec le modèle à 3 signes était supérieure à une précision aléatoire (43% vs 33%).

TABLE 3.2 – Matrice utilisée pour les calculs de précision des prédictions pour le modèle incluant uniquement les signes + et -.

Prédictions	Signe observé	
	+	-
+	1	0
-	0	1
CHANGE	0.5	0.5

3.2.2 Génération du graphe

Pour construire le modèle de régulation, nous avons utilisé la version de 2012 de la base de données PID-NCI (*Pathway Interaction Database*) [120] et l'avons récupérée au format PID-XML. Cette base ayant déjà été utilisée dans des approches de coloration de graphes sur des données humaines [68], cela nous permettait de supposer que cette base de données était adaptée pour les approches de coloration de graphes. La base entière, transformée en graphe, contient 17932 nœuds (protéines, complexes, gènes, transcription ou phénomènes de modification de protéines) et 27976 arcs (activation ou inhibition).

Afin d'orienter notre analyse, et afin de réduire la complexité de celle-ci, nous avons construit à partir de cette base un sous-graphe en connectant 3 voies de signalisation (IL6/IL6-R, IGF1/IGF1-

R et CD40), connues pour être impliquées dans le MM [69] aux gènes variants (sur/sous-exprimés) par les chemins les plus courts. Ce sous-graphe orienté et pouvant contenir des boucles, a ensuite été réduit en supprimant les nœuds avec un prédécesseur ou un successeur. Cette réduction, déjà utilisée [116], permet de réduire la complexité de l'analyse tout en maintenant les dépendances entre les nœuds du graphe.

Notons que nous avons aussi cherché à réduire le graphe initial en cherchant les gènes non-exprimés, afin de supprimer les nœuds associés. Cette approche nécessite une normalisation particulière des données des puces, la *Frozen robust multiarray analysis* (fRMA) [89] et permet, en se basant sur des données d'expression de références d'identifier les gènes dont l'expression est suffisamment forte pour qu'ils soient considérés comme exprimés [90]. Bien que potentiellement intéressante, cette méthode n'a pas été concluante pour nos analyses. En effet, celle-ci résultait en la suppression de nombreux gènes codant pour des protéines d'intérêt, en particulier des facteurs de transcription connus comme actifs dans les plasmocytes ainsi que dans les cellules cancéreuses. Il semblerait que cette dichotomie entre l'expression des gènes et l'activité de ces facteurs soit liée à la régulation qui se fait principalement au niveau post-traductionnel, amenant un niveau d'expression très faible et indépendant de l'activité de ces protéines.

Nous avons ainsi pu obtenir un premier graphe de 2269 nœuds, 2683 arcs et connectant 529 gènes variants. La réduction du graphe a permis de récupérer un graphe de 596 nœuds et 960 arcs (Figure 3.2) contenant 529 gènes observés et 67 nœuds non observés dont 23 protéines, 33 complexes, 2 processus biologiques et 9 réactions de protéines.

Nous avons ensuite pu analyser la cohérence graphe-données, en confrontant ce graphe avec les 611 séries de données discrétisées pour prédire l'état des autres nœuds du graphe (Tableau 3.3) avec une correction de type MCOS (voir section 2.5.5).

TABLE 3.3 – Répartition des observations et prédictions entre les MC et NPC. Les observations sont les données extraites des profils d'expression de gènes. Les prédictions sont les données prédites après confrontation entre le graphe et les données observées. La dernière ligne indique les observations et prédictions totales sur l'intégralité des profils.

Signe	Observations		Prédictions	
	NPC	MC	NPC	MC
+	34 %	38 %	30 %	31 %
-	34 %	51 %	29 %	36 %
0	32 %	11 %	14 %	3 %
change	—	—	7 %	6 %
Not+	—	—	2 %	1 %
Not-	—	—	3 %	1 %
?	—	—	15 %	22 %
Total	2085	210975	3279	153181

3.3 Analyse des prédictions

3.3.1 Mise en forme des prédictions

Dans cette analyse, nous avons cherché à comparer les prédictions des individus malades (MC) et normaux (NPC). Pour cela, pour chaque nœud et pour chaque individu, nous avons décomposé les prédictions en 3 valeurs booléennes. Ces valeurs booléennes permettent de représenter si un couple (i, s) ou $s \in \{+, -, 0\}$ fait partie de la classe de prédiction associée pour le nœud i (Tableau 3.4). Cette méthode permet de prendre en compte le caractère multivalué des classes de projection.

TABLE 3.4 – Matrice d’association entre chaque projection de prédiction et les couples (nœud, signe) utilisés pour l’analyse des prédictions.

Prédictions	Couple (nœud,signe)		
	(nœud,+)	(nœud,-)	(nœud,0)
+	Vrai	Faux	Faux
-	Faux	Vrai	Faux
0	Faux	Faux	Vrai
CHANGE	Vrai	Vrai	Faux
Not+	Faux	Vrai	Vrai
Not-	Vrai	Faux	Vrai
?	Vrai	Vrai	Vrai

3.3.2 Validation des prédictions

La confrontation entre le graphe et les données par la méthode de coloration des graphes nous a permis de prédire l’état des nœuds pour chaque ensemble de données. Afin de valider ces prédictions, nous avons comparé la précision de celles-ci avec celle obtenue avec des données randomisées. Pour cela, nous avons, pour chaque échantillon (MC et NPC), utilisé 50% des gènes mesurés $\{+,-,0\}$ pour prédire l’autre moitié, et répété ce calcul 1000 fois. Nous avons alors obtenu 2 vecteurs de précisions (Figure 3.3) en utilisant la même matrice que précédemment (Tableau 3.1). Un test de Student unilatéral entre ces 2 vecteurs nous permet de conclure que la précision de notre méthode a une précision plus importante que celle obtenue avec des données randomisées (p-valeur inférieure à $2.2e-16$). Il est à noter que la précision obtenue de 43% peut sembler faible. Cette valeur est à temporeriser en prenant en compte la matrice de précision utilisée (par exemple un nœud observé "+" et prédit "CHANGE" aura localement une précision de 0.5), ainsi que le fait que notre modèle utilise 3 signes.

3.3.3 Analyse des prédictions : MM vs NPC

3.3.3.1 Réduction du nombre de couples

Dans le but de réduire le nombre de variables, nous n'avons travaillé que sur les nœuds jamais observés, et avons exclu les booléens se référant aux couples invariants (nœuds associés à "0"). De cette manière, nous pouvons représenter les prédictions obtenues pour tous les GEP par une matrice de booléens M , de taille $2m \times (N^{MC} + N^{NPC})$; où m représente le nombre de nœuds en G mais jamais observés et N^{MC} (respectivement N^{NPC}), le nombre d'individus MC (respectivement NPC). Ainsi, M_{ij} fera référence à la prédiction du nœud i pour le profil j . Notons que M_{ij} peut être séparée en M_{ij}^+ et M_{ij}^- tel que $M_{ij}^s = \text{VRAI}$ explicite que la classe de prédiction du nœud i contient le signe s dans le profil j .

Afin d'identifier des marqueurs spécifiques aux MC, nous avons analysé M et cherché des prédictions sur-représentées entre les MC et les NPC. Pour cela, nous avons utilisé 2 approches. Une première basée sur les arbres de décision [111] et la seconde utilisant une classification basée sur la fréquence de chaque couple.

3.3.3.2 Arbre de décision

Dans le cas de la première méthode, visant à identifier les combinaisons de couples les plus spécifiques des MC, afin de contrebalancer le déséquilibre des effectifs (602 MC et 9 NPC), nous avons multiplié le poids des NPC par 67. L'arbre de décision obtenu (Figure 3.4) nous montre que la combinaison des couples (JUN/FOS[n,-] et FOXM1*[c,-] est associée à la majorité des MC (73%) et peut discriminer efficacement les MC des NPC. Notons que le nœud JUN/FOS[n] représente un complexe protéique, composé de JUN et FOS, localisé dans le noyau, tandis que FOXM1*[c] représente la protéine FOXM1, phosphorylée et localisée dans le cytoplasme (la syntaxe complète des nœuds est donnée dans la figure 3.2. De plus, nous pouvons identifier un autre groupe de MC (13%), caractérisé par la présence de (JUN/FOS[n,-] et l'absence de (FOXM1*[c,-] et (SRC*,-)

3.3.3.3 Analyse par fréquence

Pour l'approche par fréquence, nous avons calculé pour chaque groupe (MC et NPC) et pour chaque couple (i, s) , $i \in V$ et $s \in \{+, -\}$ le score de fréquence par la formule suivante :

$$FS_{i,s}^C = \frac{1}{N^C} \sum_{j=1}^{N^C} M_{ij}^s \quad (3.1)$$

où C représente la classe MC ou NPC, N^C représente les effectifs de la classe C , M_{ij}^s représente la valeur booléenne du couple (i, s) pour l'individu j , et s le signe $\{+, -\}$ assigné au nœud i . Ensuite, nous avons trié les résultats en se basant sur la p-valeur d'un test de Fisher entre les scores de fréquence entre NPC et MC pour déterminer les couples les plus spécifiques des MC. Dans le tableau 3.5, nous montrons les 5 meilleurs couples avec $FS_{MC} > FS_{NPC}$. Pour chacun de ces couples, nous avons vérifié le nombre de gènes (sur les 529) connectés à partir du nœud concerné (colonne "Connectivité" de la figure 3.5). Dans la figure 3.2, nous montrons (nœuds

en jaune, avec label en gras) comment les 5 protéines ou complexes identifiés dans l'approche par fréquence sont connectés selon les informations de PID-NCI.

TABLE 3.5 – 5 meilleurs résultats pour l'analyse par fréquence. FS^{NPC} et FS^{MC} indiquent le score de fréquence respectivement pour les NPC et les MC. La colonne "Référence" liste les publications en accord avec le signe prédit. La colonne "Connectivité" renvoie à la proportion de gènes connectés à partir du nœud prédit. La colonne "Observation" montre la répartition des observations pour les gènes associés au nœud sans considérer les informations du graphe. Dans le cas de JUN/FOS[n], c'est le gène *FOS* qui est pris en compte dans la colonne "Observations"

Nœud prédit	Signe	FS^{NPC}	FS^{MC}	p.val (Fisher)	Référence	Connectivité	Observations	
							+	-
JUN/FOS[n]	-	0.444	0.956	2.65E-005	[109, 142, 117, 24, 40]	8/529	373	137
FOXM1*[c]	-	0.222	0.774	7.97E-004	[133, 50]	529/529	85	265
STAT6*[c]	-	0.222	0.764	1.05E-003	∅	8/529	30	429
EGF/EGFR*[m]	+	0.556	0.935	2.08E-003	[85, 84, 62]	529/529	79	4
Src*	+	0.556	0.935	2.08E-003	[55, 29, 60]	529/529	110	48

Nous pouvons observer que l'inhibition du complexe JUN/FOS[n] (syntaxe expliquée à la figure 3.2) est prédite pour 95% des MC, contre 44% des NPC, en faisant le couple le plus efficace pour distinguer les MC des NPC. FOXM1*[c] et STAT6*[c] sont prédits comme ayant un niveau d'activité réduite chez les MC. Les protéines FOXM1*[c] et STAT6*[c] sont prédites comme majoritairement inhibées chez les MC, on peut donc associer cette prédiction à un niveau d'activité réduite chez les patients malades. Notons que cette réduction prédite semble corrélée avec leur niveau d'expression chez respectivement 76% et 93% des MC (colonne Observations). L'approche par fréquence identifie aussi la présence de (Src*,+) comme un marqueur fort pour les MC. Ce résultat est à mettre en lien avec l'arbre de décision (figure 3.4) qui identifiait cette fois-ci l'absence de (Src*,-) comme un marqueur des MC. Ce résultat renforce l'hypothèse de la complémentarité de nos 2 approches de classification. Les 2 méthodes identifient (JUN/FOS[n,-) et (FOXM1*[c,-) comme des marqueurs importants pour les MC. De la même manière, une classification par forêts aléatoires (figure 3.5) nous a permis d'obtenir les mêmes couples.

L'activité du complexe JUN/FOS comme marqueur spécifique

Le premier couple identifié (JUN/FOS[n,-) correspond à l'inhibition du complexe composé des protéines JUN et FOS. Ces protéines forment un facteur de transcription hétérodimérique nommé AP-1. Ce facteur de transcription est connu pour jouer un rôle en tumorigenèse, en raison de son implication dans la survie et la prolifération cellulaire dans plusieurs pathologies (myélome multiple, cancer du sein). Les méthodes de classification ont montré que (JUN/FOS[n,-) était le meilleur couple pour distinguer les MC des NPC et semblent indiquer que ce complexe est inhibé chez l'énorme majorité des patients atteints du MM. Ces résultats semblent en phase avec la littérature, qui montre une réduction de l'activité du complexe AP-1 dans les cellules myélomateuses [24]. Dans les lignées de MM, ce complexe serait une manière de se protéger de l'apoptose [142]. Nous pouvons aussi noter que JUN/FOS[n] (figure 3.6) est connecté au gène de la protéine BIM (protéine pro-apoptotique) dont l'expression du gène est

identifiée comme inhibée chez 65% des MC. Enfin, ce résultat est d'autant plus intéressant qu'il est difficile d'inférer l'activité du complexe AP-1 à partir de données d'expression de JUN et FOS en raison de sa nature hétérodimérique et de son activation post-transcriptionnelle. Cette particularité illustre l'un des avantages de la méthode de coloration des graphes.

L'activité du facteur de transcription FOXM1 comme un marqueur de survie

Le second couple identifié, (FOXM1*[c],-), correspond à l'inhibition de la protéine phosphorylée FOXM1 dans le noyau. FOXM1 (Forkhead box protein M1) est un facteur de transcription connu dans le MM et étudié comme facteur pronostique et comme cible thérapeutique [50]. En raison de la réduction du graphe et de notre modèle de raisonnement, FOXM1*[c] peut être considéré comme représentatif de l'activité du facteur de transcription FOXM1. Dans un premier temps, nous avons analysé l'expression du gène *FOXM1* dans le groupe MC et nous avons pu identifier que les individus MC avec une activité de FOXM1 prédite comme inhibée avaient un niveau d'expression du gène *FOXM1* statistiquement inférieur à ceux dont l'inhibition n'est pas prédite (figure 3.7).

Il est possible de séparer les 602 patients MC analysés selon le traitement reçu. 450 font partie de la cohorte dite VD et tous ont reçu un traitement à base de Velcade-dexaméthasone puis une auto-greffe. Les 152 restants font partie de la cohorte Non-VD et ont eu uniquement l'autogreffe. Un travail précédent avait étudié le lien entre la survie globale et le niveau d'expression de *FOXM1* (figure 3.8, gauche) sur la cohorte VD. Cette étude a permis d'identifier une survie moins forte chez les individus avec l'expression de *FOXM1* la plus forte (appartenant au 4ème quartile) par rapport aux autres groupes (1er, 2nd, et 3ème quartile). Nous avons cherché s'il était possible d'établir un lien entre les prédictions de l'activité de FOXM1 et la survie globale (*overall survival* : OS) des patients. Un test de régression de Cox entre le groupe des MC avec la présence du couple (FOXM1*[c],-) et ceux dont les prédictions n'incluent pas ce couple a renvoyé une p-valeur de 0.09, ne permettant pas de conclure à une différence mais à une tendance à une meilleure survie avec une activité de FOXM1 réduite. Ce résultat, moins concluant qu'avec l'analyse de l'expression du gène de FOXM1, est à mettre en lien avec le fait que notre modèle travaille sur des expressions de gènes discrétisées et non plus continues entraînant obligatoirement une perte d'information. Ainsi, si l'analyse de l'expression de *FOXM1* permet de conclure à un lien avec la survie, l'analyse des prédictions de l'activité de FOXM1, nous permet d'identifier une tendance à une meilleure survie lorsque celle-ci est inhibée. Ces deux résultats semblent en phase avec la littérature considérant que la sur-expression de FOXM1 serait un facteur de risque dans le MM [133] et que son inhibition peut entraîner l'apoptose des cellules cancéreuses. Nous pouvons aussi signaler que dans le cas de JUN/FOS, aucun lien avec la survie n'a pu être trouvé.

3.3.4 Analyse du lien entre survie et prédictions

Pour élargir l'analyse précédente, nous avons étudié pour chaque couple s'il était possible d'établir un lien entre la survie et la présence de celui-ci dans les prédictions des 450 patients analysés précédemment avec un traitement homogène. Contrairement aux analyses précédentes, où nous

comparisons les MC aux NPC, nous avons étudié les sous-groupes de patients MC, divisés selon leurs prédictions. Pour un couple nœud-signe, nous considérons le groupe des MC ayant ce couple prédit et ceux ne l'ayant pas. Dans ce contexte, nous avons intégré les prédictions de signes invariants. Pour chaque couple de prédiction (nœuds, signe), le groupe de patients a été divisé en 2, ceux ayant cette prédiction et ceux ne l'ayant pas. Ainsi, nous pouvons comparer par une régression de Cox si ces 2 populations ont des durées de survie statistiquement différentes et donc si le couple étudié peut apporter une information pronostique. Du fait du nombre important de couples analysés (201 couples) et donc de tests réalisés, le risque important de faux-positifs (false discovery rate : FDR) a été pris en compte. L'analyse a permis d'identifier 3 couples, ayant une p-valeur inférieure à 5% et un FDR inférieur à 25% (Tableau 3.6).

TABLE 3.6 – Liste des 3 couples identifiés indépendamment pour leur impact sur la survie. La colonne "Rapport des hasards" indique l'écart entre les 2 distributions. Cette valeur indique un effet protecteur si elle est inférieure à 0, et inversement un effet délétère si supérieure à 1. La colonne "P-valeur" donne la p-valeur renvoyée par une régression de Cox. La colonne FDR donne la probabilité que ce couple soit un faux-positif.

Couple	Rapport des hasards	P-valeur	FDR
(G1/S_transition of mitotic cell cycle,+)	2.588	<0.0001	<0.0001
(G1/S_transition of mitotic cell cycle,-)	0.326	<0.0001	<0.0001
([P06400 Mod3927 :pid_m_200070 :Q14186]@nucleus,+)	0.494	0.0009	0.0599

Les 2 premiers couples identifiés sont associés au même nœud "G1/S transition of mitotic cell cycle" qui peut être vu comme un élément-clé de la prolifération cellulaire. Ces résultats associent ce nœud au signe + pour un effet délétère (figure 3.9, courbe 1), et montrent un effet protecteur quand associé au signe - (figure 3.9, courbe 2). Il est tout aussi intéressant de voir que les signes "+" et "-" semblent avoir un effet opposé chez les individus. De plus, ce nœud étant la première phase de la division cellulaire, on peut aisément comprendre qu'une tumeur ayant une activité de prolifération forte aura un effet nocif pour la survie d'un patient. Le troisième couple, ([P06400|Mod3927 :pid_m_200070 :Q14186]@nucleus,+) a pour nœud la protéine RB (codée par le gène *RBI*) sous sa forme phosphorylée et associée à la protéine TFDP1 dans le noyau. Les individus ayant ce couple prédit semblent statistiquement avoir une survie supérieure à ceux ne l'ayant pas (figure 3.9, courbe 3). Ce complexe est un marqueur de l'activité de la protéine RB1, connue pour son implication dans plusieurs cancers (rétinoblastome, ostéosarcome, cancer du sein, carcinome hépatocellulaire, etc.) [101]. Nous pouvons aussi noter d'ailleurs que RB1 est actuellement en cours d'étude comme cible thérapeutique dans le MM en ciblant les protéines CD4/CDK6, régulatrices de l'activité de cette protéine [103].

Afin de savoir si ces 3 couples apportaient la même information pronostique, une analyse multivariée a été faite permettant d'identifier que les couples (G1/S transition of mitotic cell cycle,+) et (G1/S transition of mitotic cell cycle,-) portaient la même information, laissant supposer que ces 2 couples agissent via des mécanismes proches. Ce résultat est d'autant plus intéressant que ces couples sont associés au même nœud. On peut ainsi supposer que l'activation et l'inhibition de la prolifération, bien qu'ayant des effets opposés, vont avoir des impacts sur les mêmes voies biologiques. On peut le constater aussi sur la figure 3.9, les courbes 1 et

2 étant quasiment les mêmes (avec une inversion des colorations), montrant que les individus ayant le couple (G1/S transition of mitotic cell cycle,-) prédit n'auront pas (G1/S transition of mitotic cell cycle,+) et inversement.

Parallèlement, les données cliniques des 450 patients ont été utilisées afin d'identifier celles qui apportent une information pronostique. Cette analyse a permis d'identifier 3 facteurs pronostiques statistiquement valides ayant une incidence délétère sur la survie des patients. Ces 3 facteurs sont connus dans la littérature, ce qui nous permet d'estimer que la cohorte étudiée est représentative d'une population standard : le taux de la $\beta 2$ supérieur à 5.5 (beta2_5_5) [113, 144], la translocation entre les chromosomes 4 et 14 (t4;14) [32, 63, 113] et la présence de la délétion du bras court du chromosome 17 dans au moins 60% des cellules (del17_60) [32, 95, 113].

Nous avons pu analyser ces 2 couples sélectionnés (nous n'avons pas utilisé (G1/S transition of mitotic cell cycle,+) en raison du fait que ce couple apportait la même information que (G1/S transition of mitotic cell cycle,-)) et les 3 facteurs pronostiques identifiés via une régression multivariée afin de savoir si ces 5 variables apportaient chacune de l'information indépendante. Après analyse, nous avons pu observer que les 5 variables étudiées apportaient chacune une information fiable et indépendante (Tableau 3.7). Ainsi, l'utilisation des prédictions issues de notre modèle de coloration a permis d'identifier des marqueurs apportant une information complémentaire des marqueurs cliniques classiques sur la durée de survie des patients atteints de MM.

TABLE 3.7 – Résultat de l'analyse multivariée sur les 5 variables identifiées. La P-valeur est issue d'une régression de Cox.

Variable étudiée	Rapport des hasards	P-valeur
beta2_5_5	1.526029	0.056
t4;14	2.406754	0.000
del17p_60	3.162685	0.000
(G1/S_transition of mitotic cell cycle,-)	.4662132	0.000
([P06400 Mod3927 :pid_m_200070 :Q14186]@nucleus,+)	.5760818	0.025

Nous avons enfin voulu vérifier si les 3 couples identifiés pouvaient apporter une information pronostique sur les patients restants appartenant à la cohorte Non-VD, initialement étudiée pour les facteurs pronostiques (152 individus), qui ont eu un traitement différent. Une régression de Cox univariée sur chacun de ces couples n'a pas permis d'obtenir une p-valeur statistiquement valide. Seul le couple (G1/S transition of mitotic cell cycle,+) a une p-valeur inférieure à 10% (figure 3.10) permettant de conclure à une tendance à une survie moins bonne quand la prolifération cellulaire est activée. Ce résultat, montrant les limites de notre approche est toutefois à tempérer en raison de l'hétérogénéité des individus de ce sous-groupe, ainsi que leur nombre bien plus réduit que celui d'apprentissage. Néanmoins, ceux-ci restent à approfondir sur d'autres cohortes de patients.

3.3.5 Outils et logiciels

Pour l'analyse de la cohérence des signes, nous avons utilisé l'outil *Iggy* [129], qui permet d'utiliser l'implémentation en ASP de ce modèle. La génération du graphe et la transformation des prédictions en couples nœud-signes ont été implémentées en Python 2.7 en utilisant la librairie *NetworkX* [54]. L'analyse des prédictions statistiques a été menée via le logiciel *R* [112]. Le calcul des prédictions pour les 611 profils d'expression prenait environ 5 minutes sur une machine standard.

3.4 Simulation de l'effet d'un perturbateur

Nous souhaitons aussi, durant cette thèse, étudier la possibilité d'utiliser l'approche de coloration cohérente des graphes pour simuler l'effet d'un perturbateur. En effet, bien que nous ayons montré via les analyses précédentes que ces méthodes pouvaient servir à identifier des marqueurs-clés (couples spécifiques à une population) et des facteurs pronostiques, il peut être intéressant de pouvoir aussi identifier les nœuds ayant un impact fort sur le graphe de régulation.

3.4.1 Méthode

Nous avons cherché à quantifier, *in silico*, l'impact de la perturbation d'un nœud du réseau sur le profil d'expression d'un individu.

Pour simuler la perturbation p (activation ou inhibition) d'un nœud n sur un ensemble d'observations β , nous générons un nouvel ensemble d'observations β^* contenant β ainsi que n associé au signe correspondant à la perturbation (+ pour une activation, - pour une inhibition). Nous pouvons ensuite confronter β^* avec le graphe de régulation via une correction de type SCENFIT (voir section 2.5.5). Dans ce cas, n sera considéré comme un *input*, c'est à dire que sa coloration n'aura pas à être expliquée, et ne pourra pas être changée dans l'exploration des réparations par SCENFIT. Il est alors possible de récupérer le score de réparation, qui est indicateur du niveau de conflictualité entre le nœud perturbé et les observations. Par exemple, dans la figure 3.11-2, le SCENFIT associé à l'inhibition de B sera de 2 en raison du changement de signe du gène 3. Si le SCENFIT-repair est efficace pour évaluer le nombre de conflits avec les observations générées, il est en revanche moins adapté pour étudier les conflits établis avec plusieurs profils (le nombre d'observations pouvant varier, influençant de fait le nombre de réparations). Afin de contrebalancer ce problème, nous avons mis au point une métrique permettant d'identifier la capacité d'un nœud à générer des conflits avec les observations de plusieurs ensembles de données, le *Top perturbation score* (TPS), qui se calcule selon la formule suivante :

$$TPS_{i,s}^C = \frac{1}{NC} \sum_{j=1}^{NC} f(i,s,j), \text{ où}$$

$$f(i,s,j) = \begin{cases} 1, & \text{Si } SF_{ij}^s \geq \text{top}(SF_j). \\ 0, & \text{Sinon.} \end{cases}$$

où C représente la classe MC ou NPC et s le signe $\{+, -\}$ associé au nœud i . SF_{ij}^s représente le score SCENFIT de la perturbation du nœud s avec un signe i sur le profil j . La fonction $top(SF_j)$ va calculer le score de SCENFIT qui sépare 10% des couples ayant le plus haut score avec le reste dans le profil j . $TPS_{i,s}^C$ peut donc se traduire par le pourcentage de présence d'un couple (i, s) dans les 10% meilleurs des individus de la classe C .

3.4.2 Implémentation

Ce modèle de perturbation utilise le même outil *Iggy* [129] utilisé précédemment. Les calculs des perturbations des nœuds ont été menés sur la plateforme BIRD (www.pf-bird.univ-nantes.fr) disposant de 320 nœuds et 1.3 To de RAM.

3.4.3 Impact des perturbations sur les profils d'expression

À partir de la méthode présentée ci-dessus, nous avons cherché le TPS associé aux 5 nœuds identifiés par l'approche par fréquence avec les signes + et - (Tableau 3.8). Nous avons ensuite comparé les TPS obtenus entre les MC et les NPC afin d'évaluer la significativité de ces perturbations. Nous pouvons voir que l'activation de JUN/FOS[n] fait partie des 10% meilleures perturbations chez 74.6% des MC contre seulement 22.2% des NPC. Parmi ces 74.6% de MC, tous avaient le couple (JUN/FOS[n],-) prédit. Nous pouvons rappeler qu'il a été montré que l'activation de JUN/FOS sur des lignées de cellules MM entraînait leur mort cellulaire et une inhibition de leur prolifération [24]. Ce résultat semble d'autant plus intéressant que plusieurs approches thérapeutiques visent le complexe JUN/FOS [109, 117].

La même tendance (plus de conflits chez les MC que les NPC) s'observe lorsque FOXM1*[c] est activé, mais n'est pas significatif. Néanmoins, nous pouvons tout de même noter que l'activation de FOXM1 fait partie des 10% meilleures perturbations chez 36.4% des MC et que parmi ceux-ci, 96.7% ont la prédiction (FOXM1*[c],-).

Pour les autres protéines et complexes, nous voyons que la différence entre MC et NPC n'est pas significative. Malgré tout, nous remarquons que la P-valeur associée à une perturbation s'opposant à une prédiction identifiée avec la classification par fréquence (Table 3.5) aura toujours une valeur inférieure à la P-valeur d'une perturbation dans le même sens.

3.5 Conclusion

Dans cette contribution, nous avons proposé une approche utilisant la coloration cohérente des graphes pour étudier un ensemble de 602 profils d'expression issus de patients atteints du MM comparés à 9 profils de cellules plasmocytaires saines. Notre objectif était d'identifier les mécanismes sous-jacents à ces profils en identifiant les états des protéines pouvant avoir un rôle central dans l'expression des gènes. Notre approche repose sur un raisonnement confrontant des graphes de régulation et des observations discrétisées d'expression des gènes sous forme de programmes logiques qui combinent ces informations. Celle-ci peut se résumer en 2 étapes. Premièrement, nous générons un graphe de régulation, permettant de connecter les gènes sur/sous-exprimés chez les MM à partir de 3 récepteurs cellulaires. Ensuite, nous avons confronté ce

TABLE 3.8 – Top Perturbation Score pour les nœuds identifiés par l’approche par fréquence. La colonne Dir informe sur le signe de la perturbation (+, activation and -, inhibition). TPS représente la fréquence de présence de la perturbation dans les 10% meilleures perturbations chez les profils MC (TPS^{MC}) ou NPC (TPS^{NPC}). La colonne P-val donne la P-valeur retournée par un test unilatéral de Fisher. Les valeurs en gras indiquent les couples opposés à ceux identifiés précédemment par l’analyse par fréquence (Tableau 3.5).

Node	Dir.	TPS^{NPC}	TPS^{MC}	P-val
JUN/FOS[n]	+	22.2%	74.6%	0.001
	-	44.4%	0.5%	1
FOXM1*[c]	+	11.1%	36.4%	0.107
	-	55.6%	19.1%	0.997
STAT6*[c]	+	33.3%	55.0%	0.169
	-	44.4%	21.9%	0.970
EGF/EGFR*[m]	+	0.0 %	0.3%	0.971
	-	0.0 %	3.5 %	0.728
Src*	+	0.0 %	1.3%	0.887
	-	11.1%	33.4%	0.150

graphe avec les données transcriptomiques discrétisées avec le logiciel *Iggy*, permettant de prédire les couples (*node, sign*) représentant les états des entités biologiques. Via 2 approches de classification, nous avons été capables d’identifier des couples spécifiques aux MC en comparaison des NPC. Enfin, utilisant notre modèle de coloration, nous avons étudié l’effet d’une perturbation *in silico* sur un profil d’expression.

Un avantage de cette méthode est la possibilité d’inférer l’état d’une protéine à partir de données transcriptomiques, en utilisant le caractère causal des interactions de PID. Cela peut être intéressant pour des modèles biologiques, et plus particulièrement dans la modélisation du cancer où les données transcriptomiques sont plus facilement accessibles. Après avoir vérifié statistiquement la qualité de nos prédictions, nous avons pu identifier 5 entités biologiques (1 complexe et 4 protéines), chacune associée à un signe dont la combinaison est spécifique aux MC et avons pu trouver que le complexe AP-1 et le facteur de transcription FOXM1 sont inactivés dans la majorité des individus atteints du MM. Enfin, nous avons pu identifier une tendance chez les individus pour lesquels nous prédisions l’inactivation de FOXM1 à avoir une survie plus importante. De la même manière, nous avons pu associer l’activation de la prolifération cellulaire et l’inhibition de la protéine RB1 comme facteurs de mauvais pronostic. Ces facteurs étaient déjà connus dans la littérature. Ces résultats nous permettent de valider les prédictions de notre approche, et montrent qu’il est possible de les utiliser pour des objectifs de classification inter-groupe, ainsi que pour caractériser des sous-groupes différenciés par leur survie.

Nos résultats sur les perturbations *in silico* d’un système sont aussi encourageants bien qu’à tempérer. Dans notre cas, nous avons travaillé uniquement sur des perturbations simples, et une possibilité de poursuite de ces travaux serait d’étudier les combinaisons de perturbations.

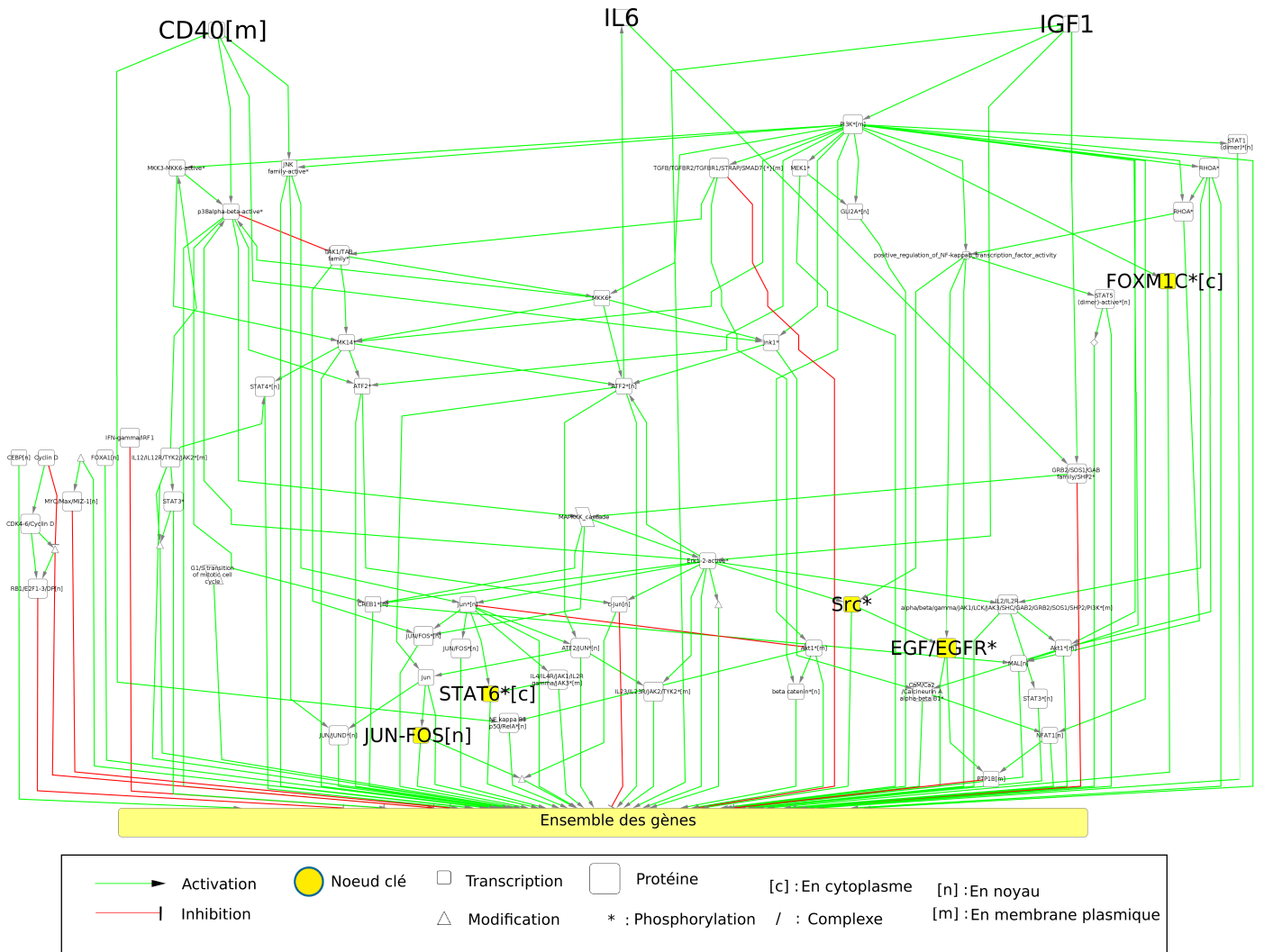


FIGURE 3.2 – Représentation du sous-graphe obtenu à partir de PID-NCI. CD40, IL6 et IGF1 (nœuds en haut) sont les 3 voies étudiées. Les 529 gènes variants ont été fusionnés dans cette représentation en un nœud “Ensemble des gènes”. Les 5 meilleurs nœuds sur la base de leur score de fréquence sont indiqués en gras, et colorés en jaune. Nous utiliserons la même syntaxe des nœuds pour le reste de ce document. Notons aussi que dans cette représentation, les arcs des gènes (“Ensemble des gènes”) aux protéines ont été supprimés pour plus de clarté.

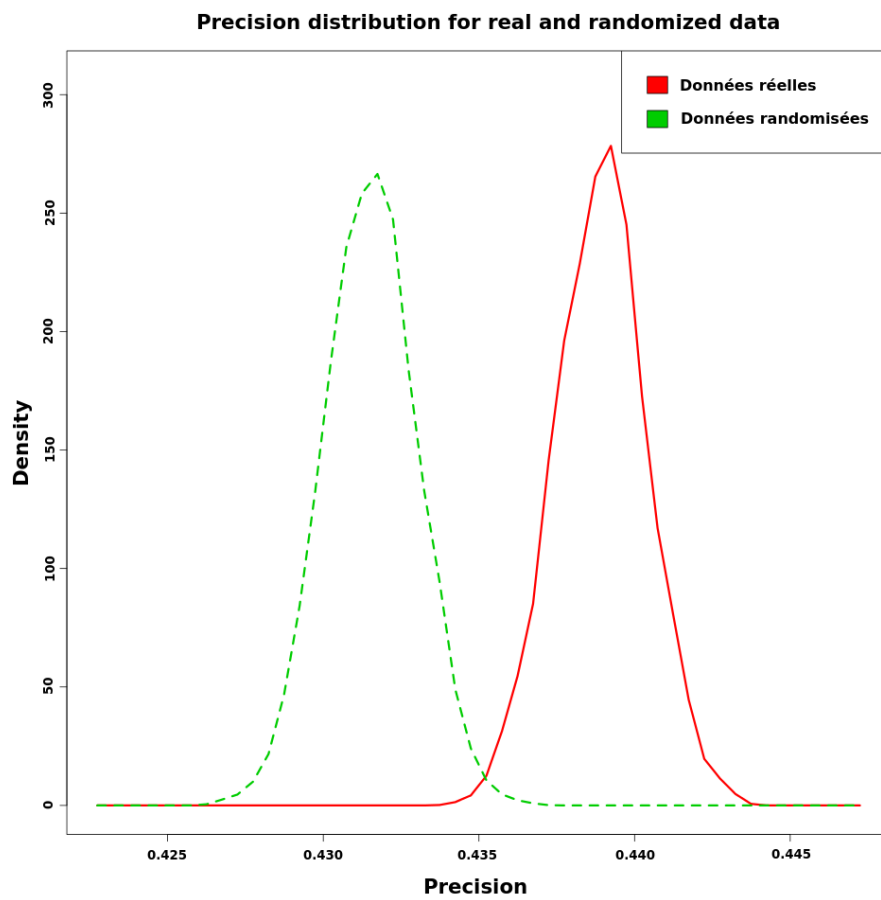


FIGURE 3.3 – Distribution des précisions des prédictions par coloration cohérente des graphes avec données réelles et données randomisées.

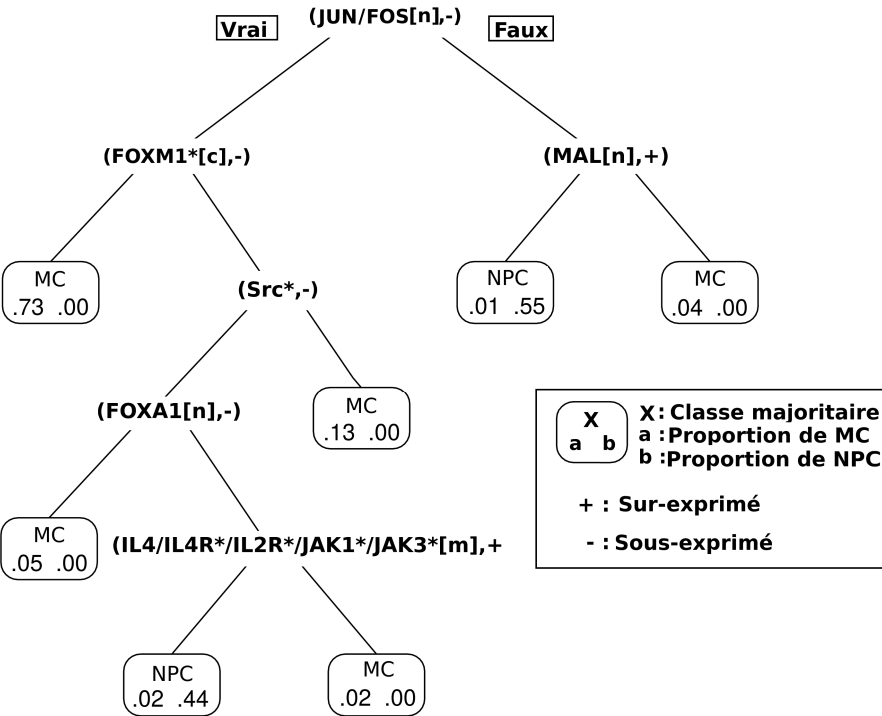


FIGURE 3.4 – Arbre de décision basé sur les signes prédits des nœuds.

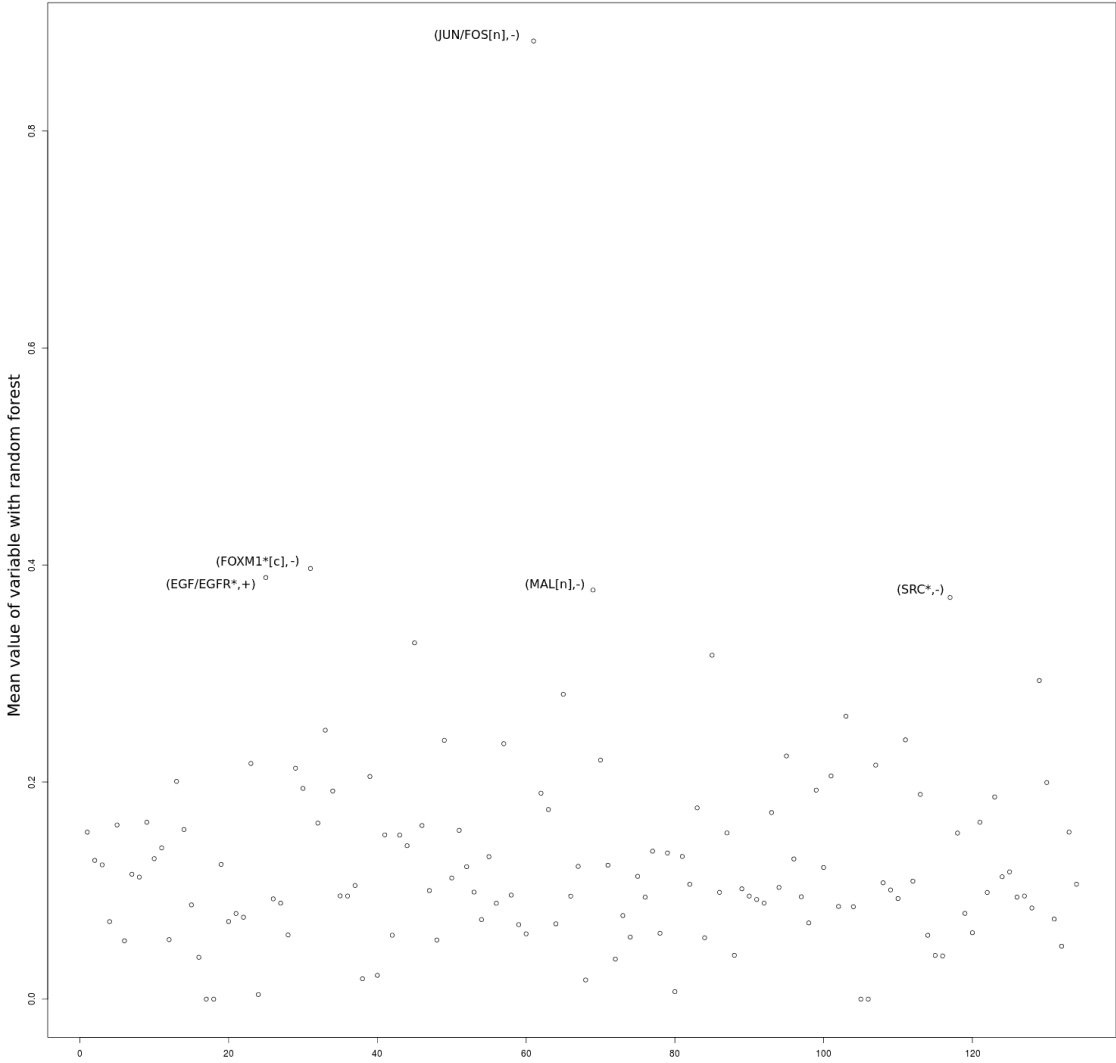


FIGURE 3.5 – Forêt aléatoire basée sur les signes prédits des nœuds.

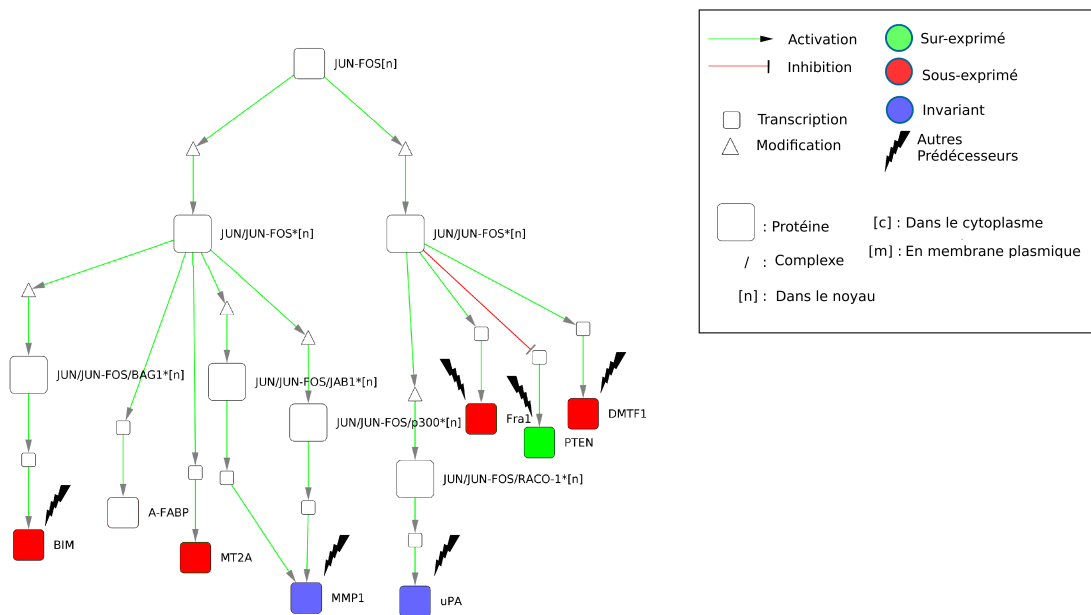


FIGURE 3.6 – Sous-graphe de JUN/FOS[n] vers les gènes variants avec un exemple d’observations issues d’un profil MC. Le graphe est le même pour chaque individu tandis que la coloration sera spécifique à chacun. Les nœuds associés à un éclair ont d’autres prédécesseurs que JUN/FOS[n] et donc peuvent voir leur coloration expliquée par d’autres nœuds.

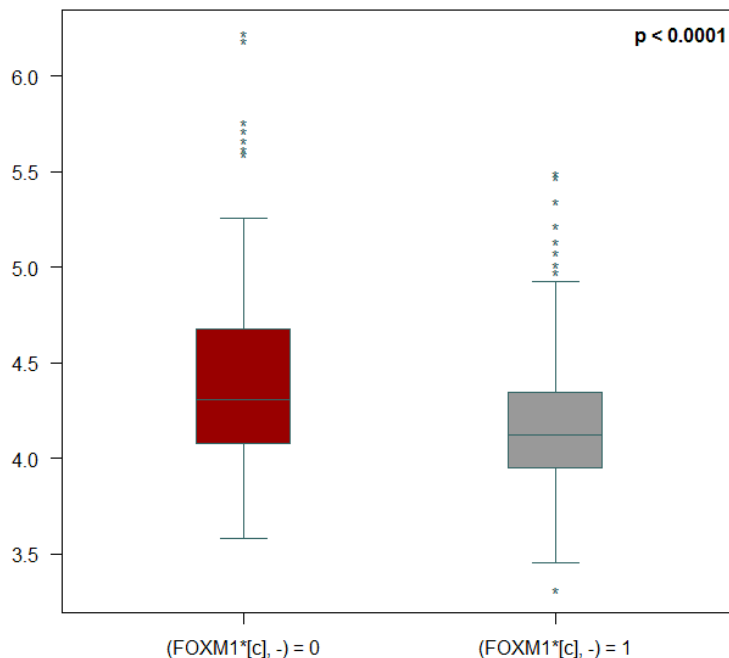


FIGURE 3.7 – Comparaison des niveaux d’expression du gène *FOXM1* entre les individus MC dont l’inhibition de FOXM1 n’est pas prédite (gauche) et les MC dont l’inhibition est prédite (droite).

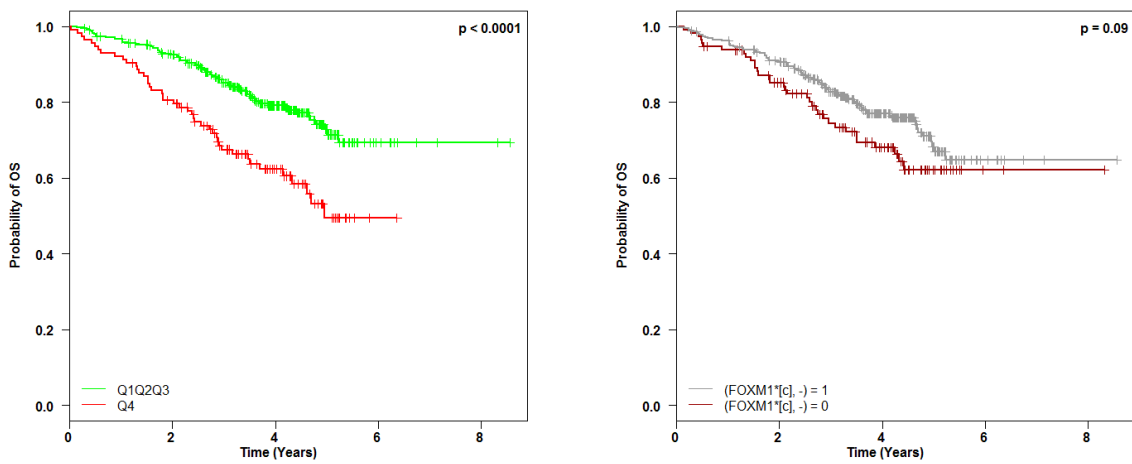


FIGURE 3.8 – Gauche : courbe de survie des individus MC selon le niveau d’expression du gène *FOXM1*, Q1Q2Q3 représente les 3 premiers quartiles d’expression du gène *FOXM1* et Q4 le dernier quartile. Droite : courbe de survie des individus MC selon la prédiction du couple (FOXM1*[n,-])

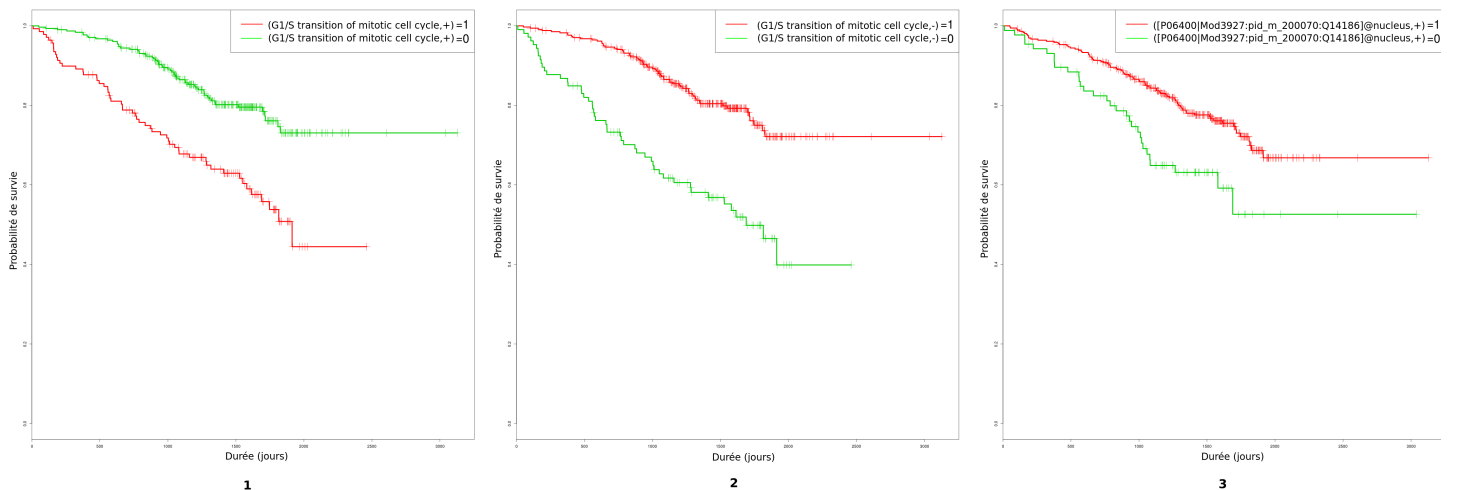


FIGURE 3.9 – Courbes de survie des 3 couples identifiés pour leur impact sur la survie.

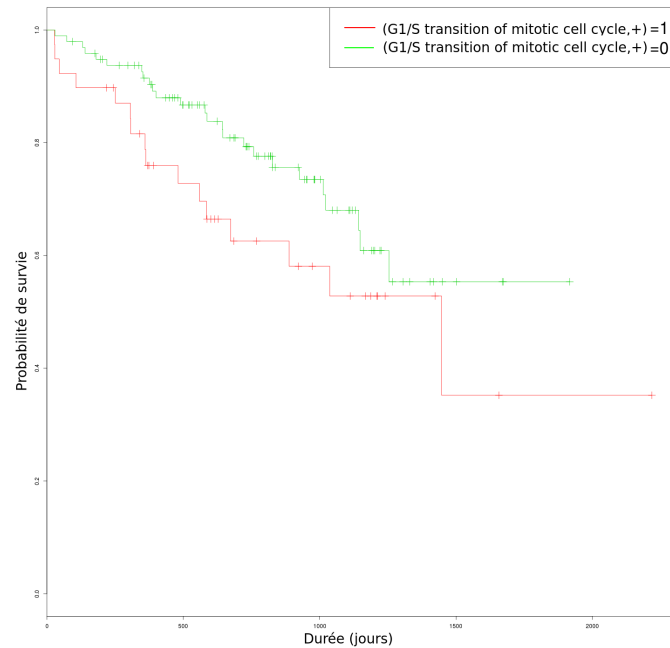


FIGURE 3.10 – Courbes de survie des 152 patients de la cohorte Non-VD selon la prédiction du couple (G1/S transition of mitotic cell cycle,+).

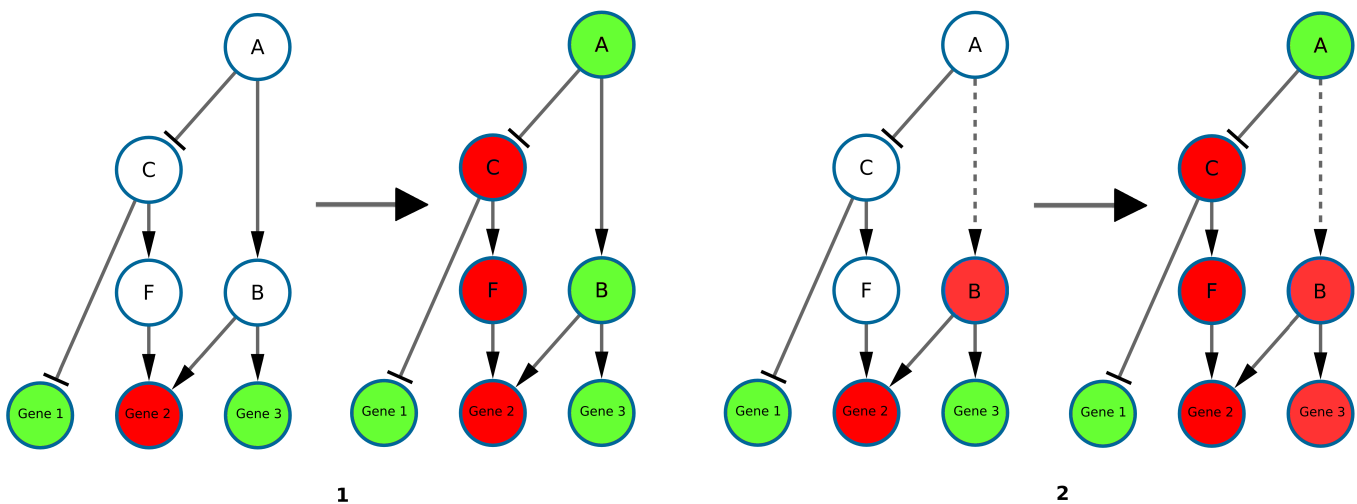


FIGURE 3.11 – 1 : Exemple de coloration sans réparation à partir d’observations. 2 : Exemple de perturbation avec l’inhibition de B, entraînant un score de SCENFIT de 2 (changement de la coloration du gène 3).



4

Exploration des colorations parfaites

Ce chapitre sera consacré à la seconde contribution de cette thèse : la recherche des colorations parfaites. Dans un premier temps, nous introduirons le cheminement ayant amené à cette contribution, ainsi que les approches n'ayant pas donné de résultat. Ensuite, nous introduirons la méthode proposée ainsi que les concepts associés. Enfin, nous présenterons deux cas d'applications. Le premier sur un exemple théorique afin de montrer le fonctionnement de cette méthode. Le second exemple portera sur les mêmes données (graphe et profils d'expression) utilisées dans le chapitre précédent afin d'illustrer les apports potentiels de cette approche.

4.1 Introduction

4.1.1 Vers la coloration fréquentielle

Nous avons montré dans le chapitre précédent les possibilités d'utilisation des colorations cohérentes de graphes pour inférer l'état des éléments biologiques non-observés. Pour la suite de cette thèse, nous avons, initialement, souhaité chercher s'il était possible d'identifier les fréquences de colorations. En effet, la méthode actuelle permet d'associer un booléen à chaque couple signe-nœud afin de savoir s'il existe une solution où le nœud est coloré de ce signe. Aussi, dans le cas où un nœud serait coloré dans 99% des colorations à "+" et "-" dans le 1% restant, Iggy prédirait ce nœud avec une projection de "CHANGE".

4.1.2 L'énumération des solutions

La manière la plus intuitive pour avoir accès à ces fréquences aurait été d'énumérer toutes les colorations possibles à partir d'un jeu d'observations et d'un réseau. Pour cette recherche de colorations, nous avons travaillé sur un modèle à 2 signes (+ et -). Cette approche est bien entendu difficile à mettre en place. Un comptage des colorations cohérentes possibles sur le graphe utilisé dans le chapitre précédent et avec un seul profil d'expression après 3 jours de calcul avait

identifié plus de 15 millions de colorations possibles (calcul arrêté car trop long). Il est à noter que l'espace des solutions à explorer était supérieur à 10^{179} . Nous avons alors décidé de chercher à travailler sur un sous-ensemble de ces colorations. Initialement, nous avons cherché à utiliser les premières colorations identifiées par *Iggy*. Néanmoins du fait de son implémentation en ASP qui explore les solutions de façon incrémentale, les colorations identifiées offraient très peu de variabilité, et donc ne pouvaient servir à caractériser l'intégralité des solutions cohérentes.

4.1.3 Vers l'identification de sous-solutions par ajout de contraintes

Une autre solution abordée aura été d'inverser le problème et de chercher non plus les solutions les plus cohérentes, mais les plus incohérentes (figure 4.1). Cette idée résultait du fait que dans un modèle à 2 signes, pour un nœud n et ayant p prédécesseurs (nœuds ayant un arc de celui-ci vers n), il n'y aura que 2 configurations incohérentes tandis que le nombre de configurations cohérentes sera de $(2^{p+1} - 2)$. Ainsi, il pouvait être intéressant de travailler sur ces solutions maximisant l'incohérence.

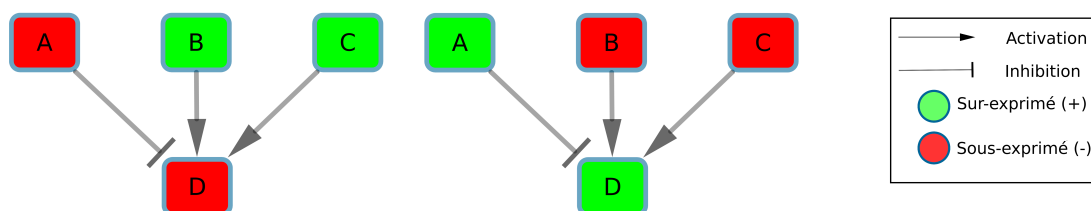


FIGURE 4.1 – Exemple de colorations incohérentes dans un modèle à 2 signes. Toutes les autres colorations possibles seront cohérentes. Nous pouvons noter que les 2 colorations sont symétriques

Deux problèmes se sont néanmoins posés sur cette approche. Bien que permettant d'accéder potentiellement à un ensemble complet de sous-solutions, il était difficile de lier les colorations maximisant les incohérences avec les observations. De plus, si cette méthode restait efficace sur des petits exemples, l'application sur le graphe utilisé dans le chapitre précédent s'est avérée non-concluante. Néanmoins, nous sommes restés sur cette idée de chercher des sous-solutions par ajout de contraintes.

4.1.4 Les solutions parfaites

Dans l'approche présentée ici, nous proposons une extension des méthodes de coloration cohérente des graphes afin d'identifier les *colorations parfaites*. Nous entendons par cela considérer les configurations de colorations en maximisant la cohérence entre les états moléculaires (sur/sous-exprimé), l'orientation des interactions ainsi que leur type (activation/inhibition). Nous avons basé ce modèle sous l'hypothèse d'une redondance des influences dans les réseaux de régulation, c'est à dire que pour un nœud donné, nous cherchons à ce que la coloration soit expliquée par un nombre maximal de prédécesseurs (figure 4.2). Nous nous sommes ensuite basés sur ces colorations parfaites pour identifier des sous-ensemble corrélés de ce graphe ap-

pelés *composants*. Ce type de méthode s’approche des identifications de modules utilisés dans d’autres approches [7, 36, 115, 39].

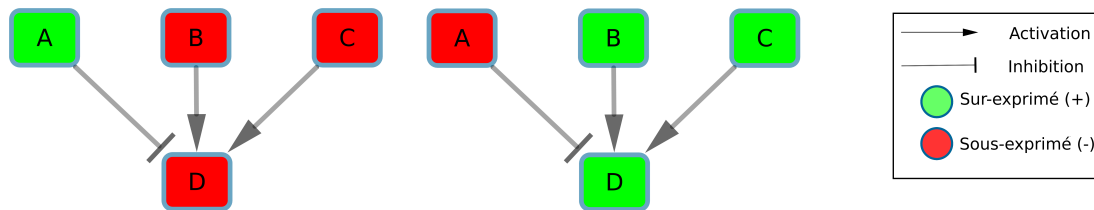


FIGURE 4.2 – Exemple des colorations parfaites dans un modèle à 2 signes. Toutes les autres colorations seront imparfaites. Nous pouvons aussi noter que les 2 configurations sont exactement symétriques

En intégrant des données d’observations, il est possible d’identifier des *composants* d’intérêt sur la base d’une métrique appelée similarité maximale, évaluant l’écart entre les observations d’un profil d’expression et les configurations de chaque *composant*. Enfin, il est possible d’associer chacun de ces *composants* à des fonctions biologiques. Nous présenterons tout d’abord la méthode, les différentes optimisations, puis un exemple. Enfin, nous montrerons une application de cette méthode sur les données et le graphe utilisés dans le chapitre précédent. Notre méthode aura permis de générer 15 *composants*. Un de ces *composants* a pu être identifié comme statistiquement spécifique aux MC par rapport aux NPC. En utilisant l’outil *PANTHER*, nous avons pu associer ce *composant* à des fonctions oncogéniques.

4.2 Méthode

Dans cette section, nous présentons les étapes du *workflow* proposé (figure 4.3). Il est important de noter que l’ordre des sections ne suit pas celui du *workflow* en raison du fait que certaines de ces étapes, en particulier la réduction de l’espace des solutions (réduction du graphe), sont basées sur des concepts qui nécessitent d’être introduits préalablement. Dans la Figure 4.3 nous montrons les entrées (réseau de régulation et données transcriptomiques) et sorties (*composants* et similarité maximale) de notre méthode.

4.2.1 Modélisation des colorations parfaites en ASP

4.2.1.1 Instanciation

Graphe : Un graphe $G(V, E)$ est composé d’un ensemble de nœuds V et d’arcs E .

Arc : Un arc (*edge*) est un tuple de 2 nœuds (source et cible), un signe (1 pour une activation, -1 pour une inhibition) et un poids.

```

1 % Arc activateur du node1 vers le node2, avec un poids de 2
2 edge (node1, node2, 1, 2) .
3 % Arc inhibiteur du node2 vers le node3, avec un poids de 3
4 edge (node2, node3, -1, 3) .

```

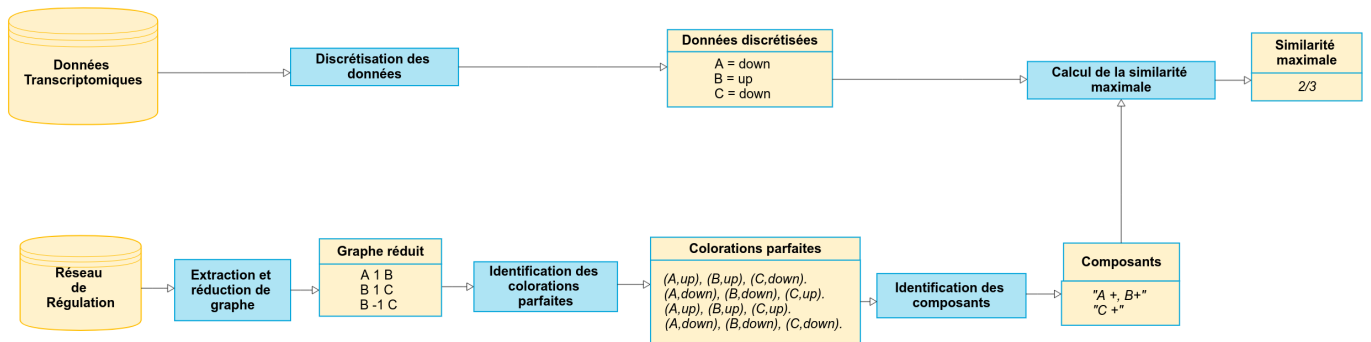



FIGURE 4.3 – *Workflow* de la méthode se basant sur les colorations parfaites afin d’identifier les *composants* à partir d’un graphe et de calculer la similarité maximale en intégrant des observations.

Nœud : Les nœuds (*node*) sont identifiés par l’union de toutes les sources et cibles des arcs.

```
5 % Nodes definition
6 node(X) ← edge(X,_,_,_).
7 node(X) ← edge(_,X,_,_).
```

Target : Une *target* est un nœud avec, au moins, un prédécesseur. Nous pouvons identifier ces *targets* en cherchant l’union de toutes les cibles dans les arcs (ligne 8)

```
8 target(X) ← edge(_,X,_,_).
```

4.2.1.2 Génération des solutions candidates

Un graphe coloré est un graphe dans lequel chaque nœud sera associé à un signe : *up* pour "+" et *down* pour "-". Ces signes font référence aux variations qui peuvent être mesurées expérimentalement pour des espèces moléculaires modélisées sous la forme de nœuds dans le graphe. La ligne 12 permet de générer toutes les possibilités de graphes colorés.

```
9 % Signs definition
10 sign(down;up).
11 % Graph coloring
12 1{coloring(I,S):sign(S)}1 ← node(I).
```

4.2.1.3 Définitions

Coloration cohérente d’un nœud. Un nœud ayant une coloration cohérente est un nœud dont la coloration est expliquée par au moins un de ses prédécesseurs dans le graphe [129]. Il existe deux possibilités pour un nœud n d’être expliqué par un prédécesseur p et selon le signe de l’arc connectant p vers n . Si l’arc est activateur (ligne 13), p devra être associé au même signe que n . Dans le cas où l’arc est inhibiteur (ligne 14), p et n devront être associés à 2 signes différents. En raison du fait qu’un nœud nécessite un prédécesseur pour avoir une coloration cohérente, cette règle ne s’applique que pour les *targets*.

```
13 consistentTarget(X) ← target(X), coloring(X,S1), coloring(Z,S2), edge(Z,X,1,_), S1=S2.
```

```
14 consistentTarget(X) ← target(X), coloring(X,S1), coloring(Z,S2), edge(Z,X,-1,_), S1!=S2.
```

Coloration imparfaite de nœud. Une coloration imparfaite de nœud est identifiée comme un nœud dont la coloration n'est pas expliquée par au moins un de ses prédécesseurs dans le graphe.

```
15 imperfectColoring(X) ← coloring(X,S1), coloring(Z,S2), edge(Z,X,1,_), S1!=S2.
```

```
16 imperfectColoring(X) ← coloring(X,S1), coloring(Z,S1), edge(Z,X,-1,_).
```

Un régulateur pondéré imparfait. Un régulateur pondéré imparfait p est un prédécesseur d'un nœud n n'expliquant pas la coloration de n . Le poids de cette règle correspond au poids de l'arc connectant p vers n .

```
17 imperfectWeightedRegulator(X, Y, 1..W) ← edge(X,Y,1,W), coloring(X,S1), coloring(Y,S2),
    S1!=S2.
```

```
18 imperfectWeightedRegulator(X, Y, 1..W) ← edge(X,Y,-1,W), coloring(X,S1), coloring(Y,S2)
    , S1=S2.
```

4.2.1.4 Contraintes d'optimisation

La méthode présentée ici identifie des colorations de graphe qui minimisent les conflits entre les sources et cibles en se basant sur les 3 définitions présentées à la section précédente. Nous proposons un exemple d'évaluation de ces contraintes dans la figure 4.4 et le tableau 4.1 de la page 66. Ainsi, nous pouvons identifier des *Colorations parfaites* en appliquant 3 minimisations successives :

Minimisation de l'incohérence. La première optimisation va sélectionner les colorations de graphes avec le nombre minimum de colorations incohérentes. Pour cela, nous allons tout d'abord identifier les *targets* incohérentes (ligne 19), puis compter la somme de ces *targets* incohérentes (ligne 20). Enfin, nous allons imposer une contrainte visant à minimiser cette somme (ligne 21).

```
19 inconsistentTarget(X) ← not consistentTarget(X), target(X).
```

```
20 sumInconsistencyTargets(X) ← X =#count{ node(Z) :inconsistent(Z) }.
```

```
21 #minimize {X@3 : sumInconsistencyTargets(X)}.
```

Minimisation des colorations imparfaites de nœuds. La seconde optimisation vise à réduire l'espace des solutions aux colorations de graphe avec le nombre minimal de colorations imparfaites de nœuds. De la même manière que précédemment, la somme des colorations imparfaites de nœuds est calculée pour chaque solution candidate (ligne 22), puis les colorations de graphes avec le nombre minimal de colorations imparfaites de nœuds seront sélectionnées (ligne 23).

```
22 sumImperfectColoring(X) ← X =#count{ node(Z) :imperfectColoring(Z) }.
```

```
23 #minimize {X@2 : sumImperfectColoring(X)}.
```

Minimisation des régulateurs pondérés imparfaits. La dernière optimisation minimise la somme des régulateurs pondérés imparfaits. Tout d'abord, pour chaque *target*, nous calculons la somme des poids de tous ses régulateurs pondérés imparfaits (ligne 24). Ensuite, nous pouvons utiliser ces sommes et calculer la somme totale des régulateurs pondérés imparfaits pour une coloration de graphe (ligne 26). Enfin, nous pouvons ajouter une contrainte, qui permettra de sélectionner les colorations de graphe avec la somme minimale des régulateurs pondérés imparfaits (ligne 27).

```

24 sumImperfectWeightedRegulatorPerTarget (X,Y) ← Y=#count{ x(A, B) :
    imperfectWeightedRegulator(A,X,B)}, imperfectColoration(X).
25 imperfectWeightedRegulatorPerTarget (X,1..W) ← sumImperfectWeightedRegulatorPerTarget (X,W
    ).
26 sumImperfectWeightedRegulator (X) ← X=#count{ x(Y,Z) :
    imperfectWeightedRegulatorPerTarget (Y,Z),imperfectColoration(Y)}.
27 #minimize {X@1 : sumImperfectWeightedRegulator (X)}.

```

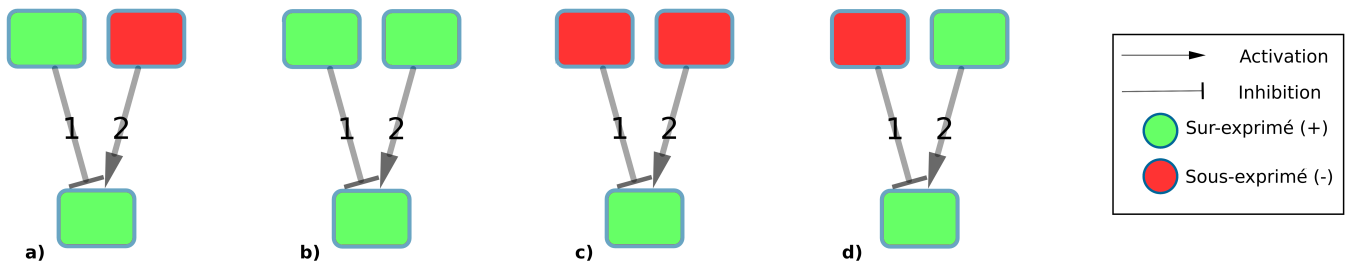


FIGURE 4.4 – Exemple de colorations possibles pour un même graphe de 3 nœuds et 2 arcs. Le label des arcs représente leur poids. L'évaluation par les contraintes est présentée dans le tableau 4.1.

TABLE 4.1 – Évaluation des contraintes sur les 4 colorations présentées à la figure 4.4. Dans le cas des recherches de colorations parfaites, c'est la coloration d qui sera gardée car minimisant les 3 contraintes.

	a	b	c	d
Coloration cohérente de la <i>target</i>	✗	✓	✓	✓
Coloration parfaite de la <i>target</i>	✗	✗	✗	✓
Régulateur pondéré imparfait	3	1	2	0

4.2.2 Identification de *composants*

Les graphes ou réseaux construits à partir de base de données, telles que NCI-PID [120] sont composés de nœuds représentant des protéines, des complexes, des gènes, des processus biologiques, ainsi que des phénomènes de transcription ou de modification protéiques. Nous définissons un *composant* comme un ensemble de nœuds qui sont coloration-dépendants, ou coloration-corrélés. Par cela, nous pouvons considérer qu'en fixant la coloration d'un nœud

dans un *composant*, les colorations des autres nœuds de ce *composant* seront fixées aussi dans les colorations parfaites. Étant donné un graphe, il est possible d'identifier l'intégralité des *composants* en construisant une matrice de corrélation pour chaque couple de nœuds à partir des colorations parfaites obtenues par le modèle présenté. Étant donné un couple de nœuds, 3 types de corrélations sont possibles (Table 4.2). Une corrélation positive telle que $b = 0$; une corrélation négative telle que $a = 0$; et une absence de corrélation telle que $ab \neq 0$. 2 nœuds étant positivement ou négativement corrélés feront partie du même *composant*.

TABLE 4.2 – Matrice de corrélation permettant de déduire les dépendances entre 2 nœuds sur les colorations parfaites. a et b représentent les occurrences de chaque combinaison de colorations.

Coloring	up	down
up	a	b
down	b	a

4.2.3 La similarité maximale

Cette étape calcule la similarité entre la coloration des *composants* et un jeu de données contenant des observations expérimentales d'un profil d'expression. En raison du modèle de colorations présenté ici et le fait que ce modèle est basé sur une coloration à 2 signes, les nœuds d'un *composant* C_i auront uniquement 2 configurations de coloration possibles. Nous les appellerons C_i^1 et C_i^2 . Nous pouvons noter aussi que C_i^1 aura la configuration inverse de C_i^2 (l'inverse d'une coloration "up" étant "down" et inversement). Nous représentons un jeu de données d'observations par un ensemble de nœuds du graphe associés à des colorations obtenues par une discrétisation préalable des mesures expérimentales. La similarité maximale (*Maximal Similarity* : MS) pour un *composant* et un jeu de données sera l'intersection maximale entre le jeu d'observations et chacune des configurations de coloration de ce *composant* divisé par le nombre de nœud observés dans le *composant* tel que :

$$MS_i = \frac{\max(|obs_i \cap C_i^1|, |obs_i \cap C_i^2|)}{|obs_i|}$$

Où i représente le *composant* analysé, obs_i les observations de nœuds dans le *composant* i . C_i^1 et C_i^2 représentent les 2 configurations de colorations pour le *composant* i .

4.2.4 Réduction de l'espace des solutions

Du fait de notre méthode de génération des solutions candidates, l'espace des solutions pour un graphe de n nœuds est de l'ordre de 2^n . En raison de notre modèle de coloration à 2 signes avec des règles symétriques, nous pouvons observer qu'une coloration de graphe et son inverse (nœuds signés "up" sont signés "down" et inversement) ont le même score vis à vis des contraintes de minimisation de l'incohérence, des colorations imparfaites et des régulateurs pondérés imparfaits. Il est alors possible d'instancier un nœud avec une coloration fixée afin de

diviser par 2 la taille de l'espace des solutions. Ainsi, par exemple avec la ligne 28, nous fixons le nœud *node0* avec la coloration "down".

```
28 coloring(node0, down) .
```

Pour améliorer cette réduction d'espace des solutions, nous proposons 3 méthodes de réduction du graphe (figure 4.5) qui peuvent être appliquées successivement sur le réseau de régulation avant de chercher les colorations parfaites. Ces méthodes identifient des nœuds du graphe qui feront partie du même *composant*. Ces nœuds seront fusionnés en un *sous-composant*. Un *sous-composant* peut être vu comme un sous-ensemble de nœuds appartenant au même *composant*, autrement dit si les nœuds d'un *sous-composant* font partie d'un seul *composant*, tous les nœuds de ce *composant* ne feront pas obligatoirement partie du *sous-composant*.

La première et seconde réductions identifient des *sous-composants*. La fusion de nœuds en *sous-composants* permet de réduire le nombre de nœuds dans le graphe et donc l'espace des solutions à explorer. La troisième méthode réduit le nombre d'arcs et peut détecter des *composants* isolés du reste du graphe.

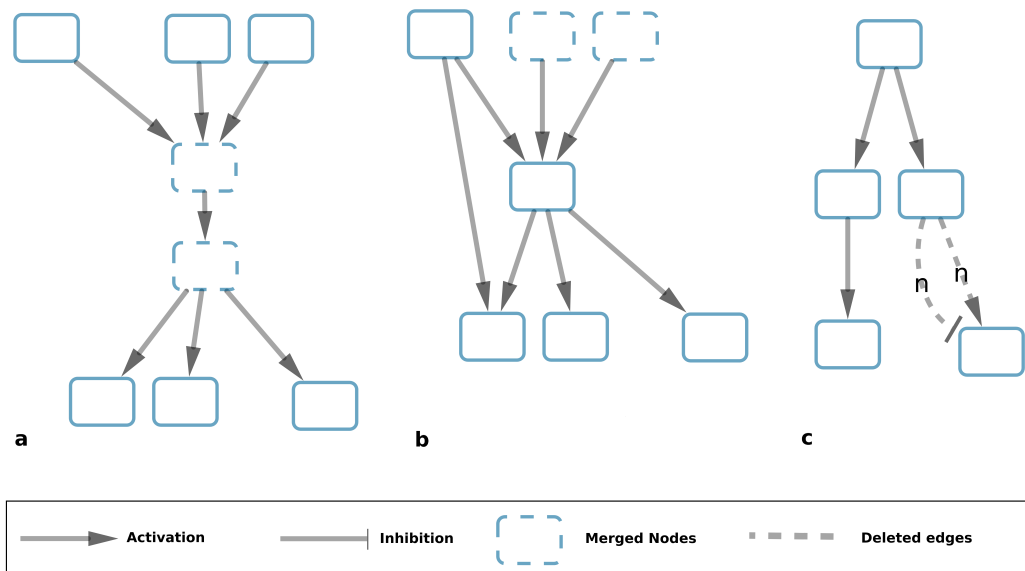


FIGURE 4.5 – Motifs recherchés par les 3 méthodes de réductions présentées dans ce chapitre. a : nœuds avec colorations corrélées dans les solutions cohérentes. b : nœuds avec colorations corrélées partageant les mêmes cibles. c : arcs avec les mêmes poids, source, cible et des signes opposés.

4.2.4.1 Réduction basée sur la cohérence (figure 4.5-a)

Cette réduction identifie des nœuds qui auront une coloration corrélée dans les solutions cohérentes. Ces nœuds sont alors fusionnés en un *sous-composant*. Cela passe par la recherche d'un motif topologique spécifique : un nœud *n* avec un seul prédécesseur *p* et un seul arc les connectant. Ce motif sera fusionné en un *sous-composant* composé de ces 2 nœuds *n* et *p* ainsi que du signe de leur corrélation dans les solutions cohérentes ("+" en cas de corrélation positive

et "-" pour une corrélation négative). Ce processus d'identification de motifs puis de fusion de nœuds sera répété jusqu'à ce qu'aucun nouveau motif puisse être détecté. Notons que fusionner un *sous-composant* avec un nœud ou un autre *sous-composant* générera un *sous-composant*.

4.2.4.2 Réduction basée sur les co-régulateurs (figure 4.5-b)

La seconde réduction identifie les nœuds qui auront une coloration corrélée dans les solutions minimisant les colorations imparfaites. Pour cela, nous cherchons un nouveau motif : 2 nœuds sans prédécesseur, ayant la même et unique cible (figure 4.5-b). Ces nœuds peuvent aussi être fusionnés en un *sous-composant*. De la même manière que précédemment, le processus de reconnaissance de motifs puis de fusion sera répété jusqu'à ce qu'aucun nouveau motif ne puisse être détecté.

4.2.4.3 Réduction basée sur l'équilibrage des arcs (figure 4.5-c)

A partir des 2 précédentes méthodes de réduction, nous obtiendrons un graphe composé de *sous-composants*, nous considérons qu'un nœud non fusionné est un *sous-composant* simple. Nous pouvons ensuite calculer le poids des arcs entre les *sous-composants* du graphe en sommant le poids de tous les arcs de même signe entre les nœuds de 2 *sous-composants*. Par ce calcul, visant à intégrer les précédentes fusions de nœuds pour le calcul des colorations parfaites, il est possible d'obtenir, entre deux *composants*, deux arcs e_1 and e_2 , de signes opposés et de poids respectifs à w_1 et w_2 . Dans ce cas, nous pouvons calculer des nouveaux poids $w_1' = w_1 - \min(w_1, w_2)$ et $w_2' = w_2 - \min(w_1, w_2)$. Dans le cas où un nouveau poids est égal à zéro (figure 4.5-c), il est possible alors de supprimer l'arc associé. Après cette réduction d'arcs, il est possible d'avoir des *sous-composants* isolés du reste du graphe car les arcs associés à ceux-ci ont eu un nouveau poids calculé à zéro. Ces *sous-composants* sont indépendants des autres *sous-composants* vis à vis de leur coloration et constitue des *composants* complets comme défini dans la section 4.2.2. Néanmoins, il sera nécessaire de stocker en information que les cibles de ces *composants* pré-identifiés seront toujours cohérents car recevant une influence positive et une influence négative du même prédécesseurs.

4.2.5 Implémentation

L'identification des colorations parfaites de graphes a été implémentée en Answer Set Programming (ASP), avec le solveur clingo 4.5.4. La réduction des graphes a été implémentée en Python 2.7 en utilisant la librairie NetworkX [54]. L'identification des *composants* à partir des colorations parfaites a été implémentée en R [112] et python 2.7. Tous les calculs présentés dans la suite de ce chapitre (extraction et réduction de graphe, identification des colorations parfaites, identifications des *composants*, calcul des MS) ont été menés sur une machine standard. Les scripts présentés ainsi qu'un exemple d'utilisation sont disponibles sur le lien suivant <https://github.com/BertrandMiannay/Iggy-POC>

4.3 Exemple

Pour illustrer la méthode qui vient d'être présentée, nous proposons un exemple d'analyse sur un graphe de 9 nœuds et 11 arcs (figure 4.6). Pour visualiser un *sous-composant*, nous le représentons par le nom des nœuds qui le composent, chacun associé au signe de leur corrélation. Ainsi, un *sous-composant* noté "A +, B -" est composé de 2 nœuds A et B qui sont négativement corrélés.

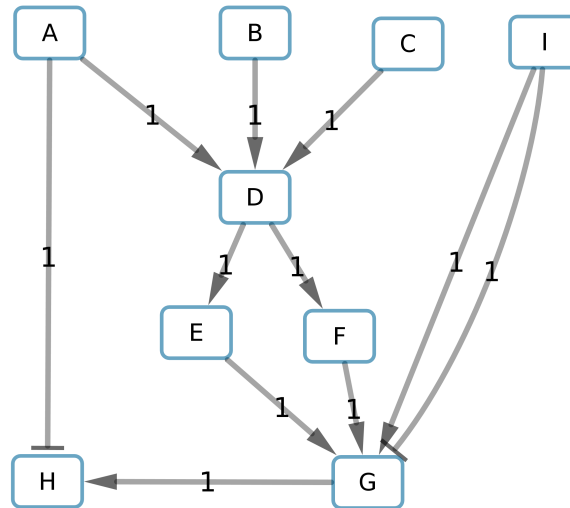


FIGURE 4.6 – Graphe utilisé pour l'exemple. Les valeurs indiquées sur les arcs indiquent le poids de ceux-ci. Les arcs "->" (respectivement "-|") représentent des activations (respectivement inhibition).

4.3.1 Réduction du graphe

Nous avons appliqué les 3 méthodes de réductions précédemment décrites pour réduire la taille du graphe. La réduction basée sur la cohérence a permis de fusionner les nœuds D, E et F (figure 4.7) car ces 3 nœuds avaient le même signe dans les solutions de coloration cohérente.

La seconde réduction basée sur les co-régulateurs a identifié les *sous-composants* "B +" et "C +" comme régulateurs du *sous-composant* "D +, E +, F +". Ces deux *sous-composants* n'ayant pas de prédécesseurs et un seul successeur partagé, ils ont pu être fusionnés en un seul *sous-composant* "B +, C +" (figure 4.8).

La dernière réduction basée sur l'équilibrage des poids des arcs identifie les arcs de "I +" vers "G +" car ayant le même poids et des signes opposés. Ces arcs peuvent donc être supprimés. Ainsi, "I +" sera isolé du reste du graphe et identifié comme un *sous-composant* indépendant du reste du graphe, et donc peut être défini comme un *composant* (figure 4.9). Nous pouvons donc sortir "I +" de la recherche des colorations parfaites et ainsi réduire l'espace des solutions. De plus, nous stockerons dans le modèle que "G +" sera cohérent et imparfait, indépendamment de ses prédécesseurs restants.

Enfin, pour réduire l'espace de solutions par deux, nous avons fixé la coloration de A à "up".

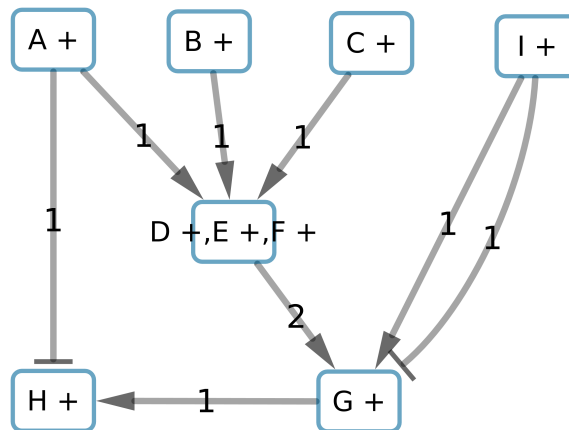


FIGURE 4.7 – Résultat de la première réduction basée sur les colorations cohérentes appliquée sur le graphe 4.6. Tous les nœuds du graphe sont des *sous-composants*.

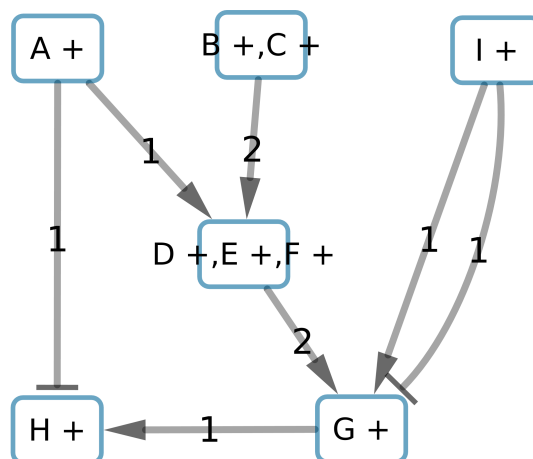


FIGURE 4.8 – Résultat de la seconde réduction basée sur les co-régulateurs, appliquée sur le graphe de la figure 4.7.

4.3.2 Colorations parfaites et identification des *composants*

Avec ce graphe réduit, nous pouvons chercher les colorations parfaites du graphe qui minimisent les colorations incohérentes, les colorations imparfaites et enfin les régulateurs pondérés imparfaits. Dans le cas de cet exemple, nous avons pu identifier 2 colorations parfaites (Table 4.3) qui minimisent les 3 contraintes présentées. Avec la coloration de "A +" fixée à "up", la recherche des colorations parfaites ne permet d'obtenir que la coloration 1. Néanmoins, la coloration 2 peut en être déduite par symétrie.

Nous pouvons observer que les *sous-composants* "A +", "B +, C +" et "D +, E +, F +" ont toujours la même coloration. De la même manière "H +" et "G +" ont toujours des colorations opposées au *sous-composants* précédents. Nous pouvons donc fusionner "A +", "B +, C +", "D +, E +, F +", "H +" et "G +" en un simple *composant*. Cette étape est réalisée en utilisant une

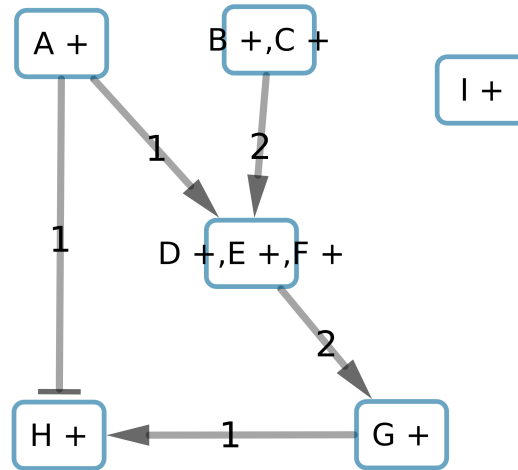


FIGURE 4.9 – Résultat de la troisième réduction de graphe basée sur l'équilibrage du poids des arcs.

TABLE 4.3 – Colorations parfaites pour le graphe réduit de l'exemple.

	A +	B +, C +	D +, E +, F +	G +	H +
Coloration 1	down	down	down	up	up
Coloration 2	up	up	up	down	down

matrice de corrélation. Au final, 2 *composants* seront identifiés et représentés dans la Figure 4.10.

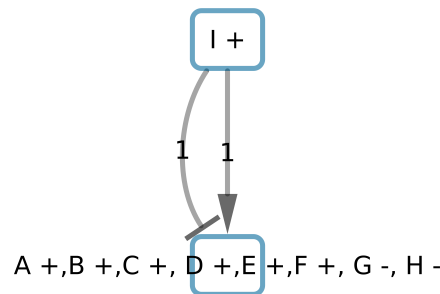


FIGURE 4.10 – Résultat de l'identification des *composants* et leurs interactions.

4.3.3 Calcul de la similarité maximale

Pour un *composant*, il existe 2 possibilités de coloration (on parle de *configuration*) en raison de la propriété de symétrie déjà utilisée précédemment. Par exemple, le *composant* "A +, B +, C +, D +, E +, F +, G -, H -" (figure 4.10) aura 2 configurations possibles :

$$C^1 = \{(A, up), (B, up), (C, up), (D, up), (E, up), (F, up), (G, down), (H, down)\}$$

$$\text{et } C^2 = \{(A, down), (B, down), (C, down), (D, down), (E, down), (F, down), (G, up), (H, up)\}.$$

Supposons un profil d'expression de gènes tel que $\{D = up, E = up, G = up\}$. Nous pouvons calculer la similarité Sim , entre le profil d'expression et chaque configuration tel que $Sim_{C^1} = 2$ et $Sim_{C^2} = 1$. La similarité maximale (MS) sera la valeur maximale entre ces 2 métriques, divisée par le nombre d'observations communes au profil et au *composant*. Ainsi, dans ce cas là, nous aurons :

$$MS = \max(Sim_{C^1}, Sim_{C^2})/3 = 2/3.$$

4.4 Application

4.4.1 Données et graphe

Dans ce chapitre, nous avons travaillé avec les profils d'expression (GEP) utilisés dans l'application précédente, à savoir 611 jeux d'expression dont 602 issus de patients atteints du myélome multiple (MC) et 9 donneurs sains (NCP). Pour chaque GEP, nous avons identifié les gènes sur/sous-exprimés par comparaison avec l'expression moyenne des NPC pour chaque gène avec un seuil à 1.2 (seuil utilisé avec la méthode précédente). De la même manière, nous avons utilisé le graphe extrait de la base NCI-PID [120] connectant 3 voies de signalisation (IL6/IL6-R, IGF1/IGF1-R and CD40) [69] vers les gènes identifiés comme sur/sous-exprimés. Ce graphe de 2269 nœuds et 2683 arcs connecte 529 gènes différentiellement exprimés. (figure 4.11, gauche). Le reste du graphe est constitué de protéines, complexes, événements de modification protéique et transcription.

4.4.2 Colorations parfaites

Nous avons réduit le graphe en utilisant les 3 méthodes présentées. La réduction basée sur la cohérence puis celle sur les co-régulateurs nous a permis de réduire le graphe à 194 *sous-composants* et 408 arcs. La réduction basée sur l'équilibrage des arcs a permis de réduire ce nouveau graphe à 193 *sous-composants*, 389 arcs et a permis d'identifier un premier *composant*. Ceci représente donc une réduction à 8% et 14% du nombre de nœuds et d'arcs par rapport au graphe d'origine. La méthode d'identification des colorations parfaites a identifié 16384 colorations pour le graphe d'origine et celui réduit (Table 4.4). Ces colorations minimisent l'incohérence à 0, les colorations imparfaites à 35 et les régulateurs pondérés imparfaits à 36. Nous pouvons noter que si le nombre de colorations est le même pour les 2 graphes, le temps de calcul, pour le graphe d'origine est plus de 300 fois supérieur à celui pour le graphe réduit. La méthode a identifié des colorations ne comportant aucune coloration incohérente. Seulement 1.5% des *targets* du graphe d'origine étaient imparfaites (coloration non-expliquée par tous les prédécesseurs). Enfin, sur les 35 *targets* colorées imparfaitement, il n'y avait un cas où le nombre de régulateurs pondérés imparfaits était de 2, les autres *targets* n'avaient qu'un seul régulateur imparfait.

TABLE 4.4 – Résultats des recherches de colorations parfaites pour le graphe d'origine et sa version réduite.

Graphe	# Nœuds	#Targets	#Arcs	Espace des solutions	Nombre de colorations incohérentes	Nombre de colorations imparfaites	Nombre de régulateurs pondérés imparfaits	Temps de calcul	Nombre de colorations
Origine	2269	2267	2683	2^{2269}	0	35	36	72 min, 12 sec	16384
Réduit	193	183	389	2^{193}	0	35	36	14 sec	16384

4.4.3 Identification des *composants*

A partir de ces 16384 colorations parfaites, nous avons pu identifier 15 *composants* (figure 4.11, droite) en utilisant la matrice de corrélation. 11 de ces *composants* étaient composés d'un simple nœud (un gène). 2 étaient composés de 2 nœuds (1 gène par *composant*). Un *composant* était composé de 422 nœuds (167 gènes) et le dernier *composant* était composé de 1832 nœuds (349 gènes).

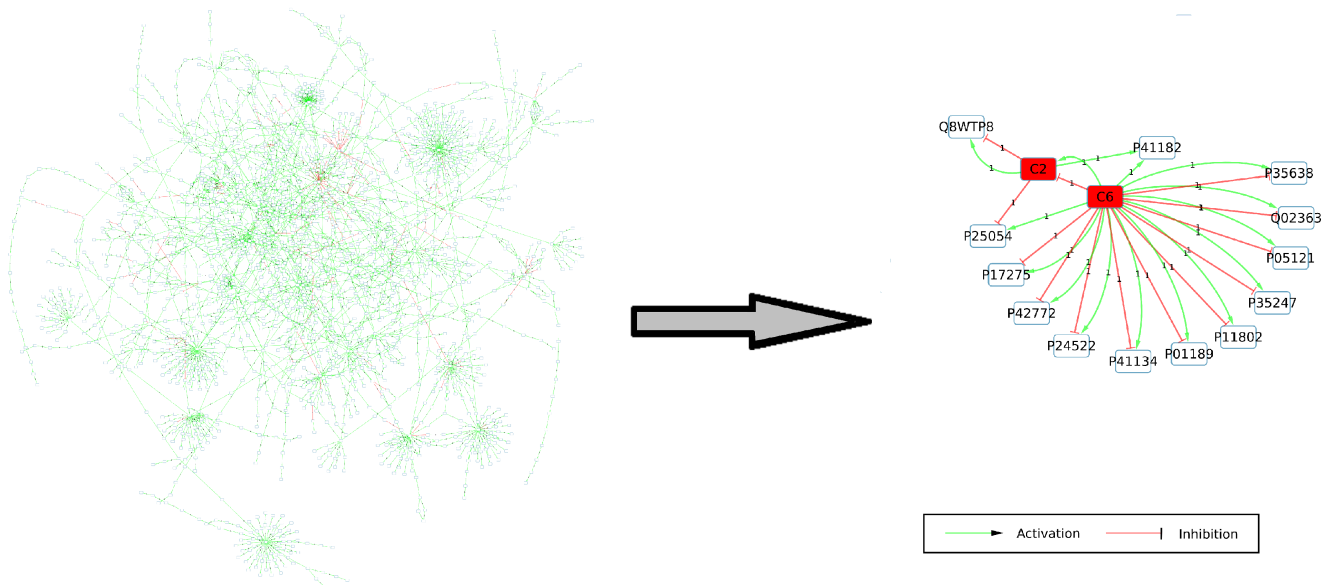


FIGURE 4.11 – Aperçu du cas d'application. Gauche : Sous graphe obtenu à partir de la base NCI-PID de 2269 noeuds et 2683 arcs. Droite : *composants* identifiés à partir des colorations parfaites. Les *composants* les plus simples sont étiquetés avec l'identifiant Uniprot correspondant au gène associé.

4.4.4 Validation des *composants*

Du fait que seulement 2 *composants* étaient composés de plus d'un gène, nous nous sommes concentrés sur ceux-ci (Table 4.5). Pour chaque profil d'expression de gène n et chaque *composant* c , nous avons calculé la similarité maximale : MS_c^n . Ainsi, nous avons pu obtenir 611

vecteurs de 2 valeurs.

Afin de valider le calcul de la similarité maximale, nous avons généré pour chaque profil d'expression 5 jeux de données randomisés en mélangeant les signes observés, de manière à ce que chaque nouveau jeu de données ait la même fréquence de signes et les mêmes nœuds observés. Comme précédemment, nous avons calculé pour chacun de ces jeux de données le MS avec les configurations des *composants*. Ensuite, pour chaque *composant*, nous avons comparé le MS entre les données réelles et les données randomisées avec un test de Welch (Table 4.5, p-valeur de validation). Les 2 *composants* identifiés avaient une p-valeur inférieure à 0.05, nous permettant de conclure à une différence significative des MS entre les données d'origine et celles randomisées.

4.4.5 Spécification des *composants*

L'étape suivante de cette analyse a consisté à identifier les *composants* spécifiques entre les NPC et MC. Pour cela, nous avons comparé le MS entre les 2 groupes de profils avec une test de Welch (Table 4.5, p-valeur de spécification). C^2 (figure 4.12) était le seul *composant* avec une p-valeur inférieure à 0.05. Nous pouvons ainsi conclure que le MS est statistiquement différent entre les MC et les NPC (figure 4.13). De plus, nous pouvons remarquer que le MS semble plus important chez les NPC que chez les MC. Or cette métrique évalue la similarité entre un état "parfait" du *composant* et les observations. On peut donc supposer que les profils MC semblent avoir des observations plus éloignées de ces configurations parfaites que les profils NPC.

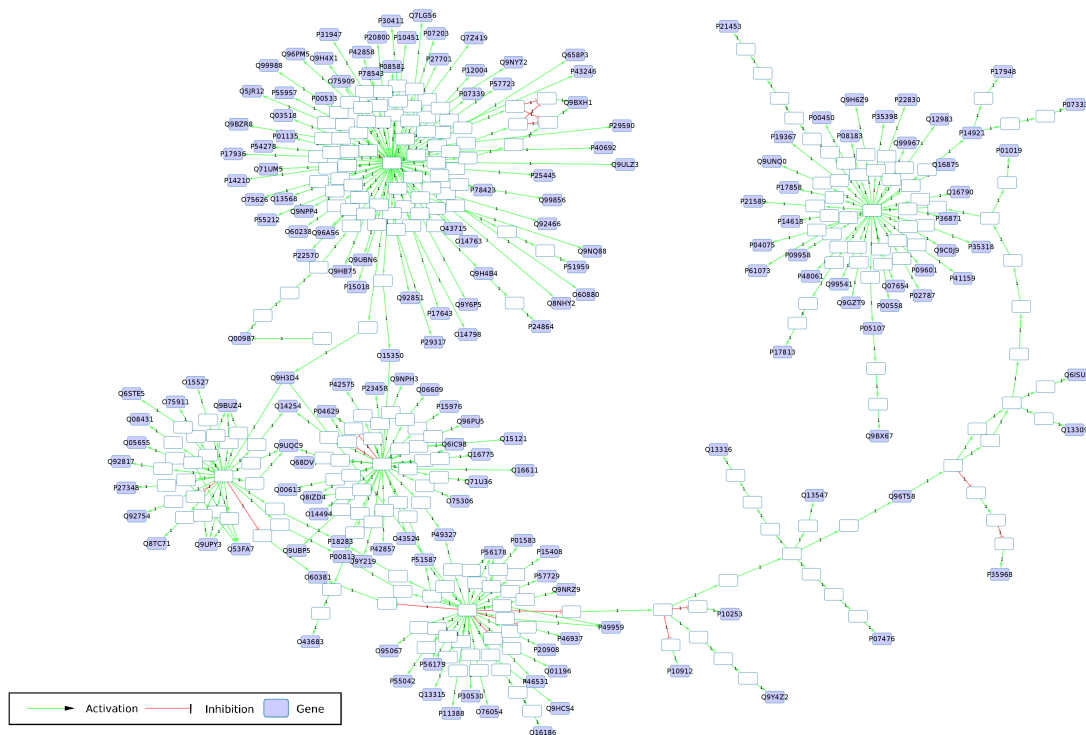


FIGURE 4.12 – Représentation du *composant 2*.

Dans le cas du *composant C⁶* (figure 4.14), la p-valeur était de 0.5725747 (figure 4.15).

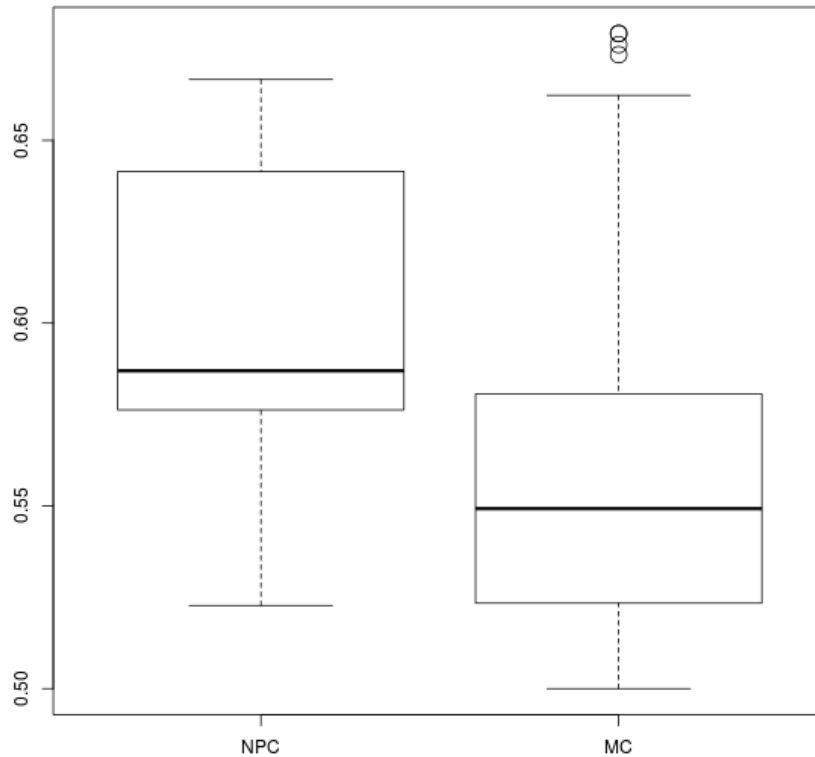


FIGURE 4.13 – Boxplot du MS calculé pour les jeux de données NPC et MC pour le *composant* 2.

u

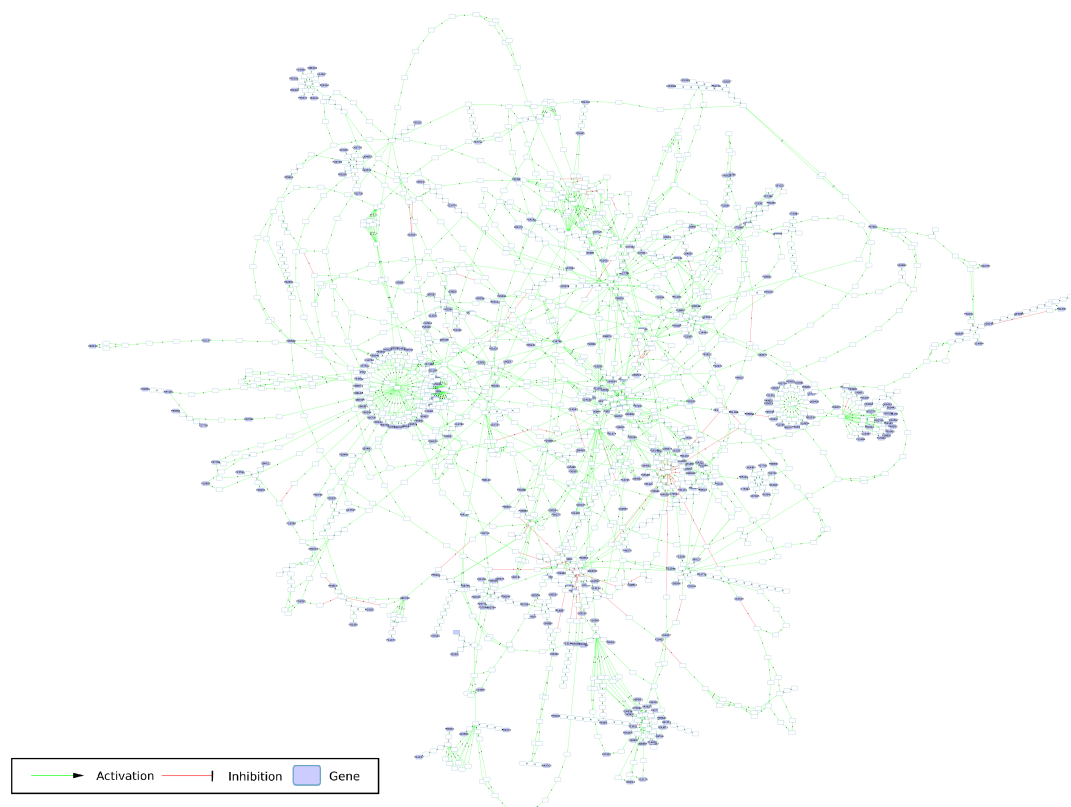
TABLE 4.5 – Résultats pour l’analyse des *composants*. La colonne “p-valeur de validation” renseigne sur la comparaison entre les données réelles et randomisées. La colonne “p-valeur de spécificité” renseigne sur la comparaison entre les données MC et NPC.

Component	# Nodes	# Genes	p-valeur de validation	p-valeur de spécificité
C^2	422	167	8.904e-03	0.019
C^6	1832	349	7.91e-05	0.573

4.4.6 Analyse biologique des résultats

Afin de pouvoir lier ces résultats d’analyse à la biologie, nous avons utilisé l’outil PANTHER [92] qui permet de réaliser des analyses d’enrichissement de gènes (*Gene ontology enrichment analysis* [127]). A partir d’un ensemble de gènes, cette analyse peut évaluer les processus biologiques sur et sous-représentés en comparaison avec un jeu de gènes aléatoires. Nous avons analysé les gènes inclus dans les *composants* C^2 et C^6 (Table 4.6 et 4.7).

Les gènes inclus dans C^2 (Table 4.6) paraissent fortement associés aux voies de la mort cellulaire, en effet, les 3 premiers processus biologiques identifiés sont liés à celle-ci. En outre, ces voies sont fortement impliquées dans les mécanismes du cancer [17]. De l’autre côté, le *com-*

FIGURE 4.14 – Représentation du *composant 6*.TABLE 4.6 – 5 premiers résultats de l’analyse d’enrichissement de gènes pour le *composant C²*. Les résultats en gras indiquent les ontologies associables à la mort cellulaire.

GO processus biologique	trouvé	attendu	Niveau d’enrichissement	P-valeur
regulation of cell death	75	11.98	6.26	6.46E-37
regulation of programmed cell death	73	11.21	6.51	8.33E-37
regulation of apoptotic process	72	11.11	6.48	4.90E-36
single-organism cellular process	149	77.70	1.92	9.90E-28
positive regulation of metabolic process	87	24.50	3.55	7.81E-26

TABLE 4.7 – 5 premiers résultats de l’analyse d’enrichissement de gènes pour le *composant C⁶*.

GO processus biologique	trouvé	attendu	Niveau d’enrichissement	P-valeur
response to organic substance	182	42.74	4.26	1.02E-68
response to chemical	203	64.12	3.17	2.13E-57
response to oxygen-containing compound	129	23.26	5.55	1.32E-56
positive regulation of biological process	233	88.29	2.64	1.39E-55
regulation of cell proliferation	132	25.67	5.14	1.98E-54

posant C⁶ (Table 4.7) ne semble pas associé à des voies redondantes. Nous pouvons néanmoins noter que l’on peut retrouver la prolifération cellulaire parmi ces voies.

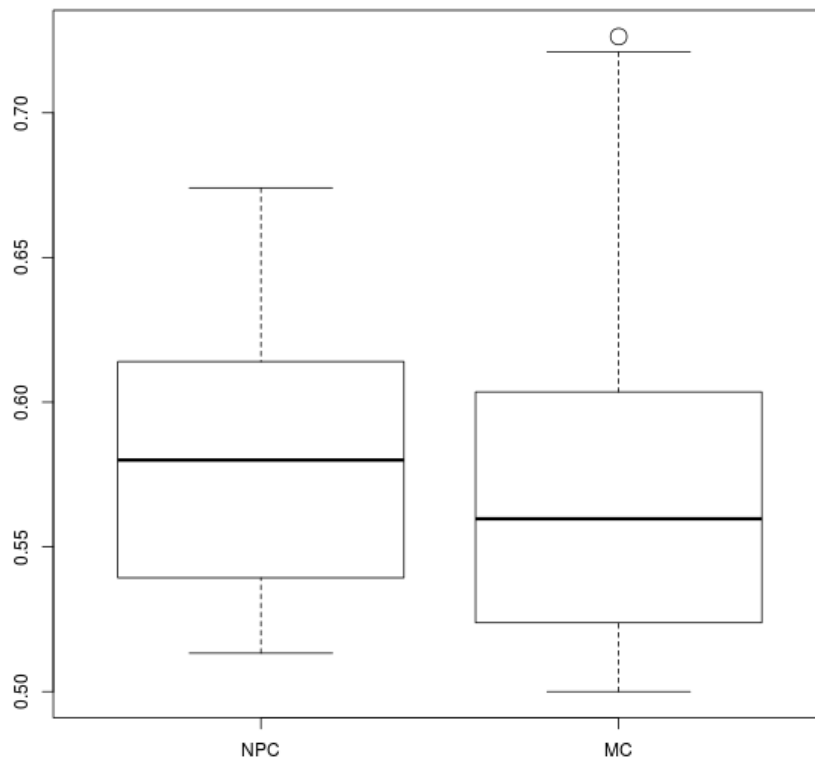


FIGURE 4.15 – Boxplot du MS calculé pour les jeux de données NPC et MC pour le *composant* 6.

4.5 Comparaison avec d'autres méthodes de classification

Afin de valider l'identification des *composants* par notre méthode, nous avons comparé les 2 *composants* identifiés par notre méthode aux *clusters* obtenus via 2 autres algorithmes de classification : ClusterOne [100] et un algorithme de *fuzzy c-means* utilisant le plug-in Cytoscape ClusterMaker [97]. Pour estimer la qualité d'un processus de classification, nous avons introduit 2 métriques. La première se base sur l'idée qu'un bon *cluster* doit avoir du sens. Nous avons donc à évaluer la significativité d'un *cluster* sur la base de son enrichissement en termes GO dits "spécifiques". En effet, les termes-GO sont organisés sous forme de graphe acyclique orienté. Le terme servant de racine est le plus général (*biological_process* : GO :0008150), plus un terme sera éloigné de celle-ci, plus celui-ci pourra être spécifique. Nous avons considéré qu'un terme-GO était spécifique si sa distance minimale de la racine était supérieure à la distance moyenne de tous les termes de la base Gene-Ontology, soit 7.5. Ainsi, pour un *cluster* i , il est possible de calculer son Enrichissement spécifique (*Specific Enrichment* : SE) via la formule :

$$SE_i = \frac{|SpecificEnrichedTerms_i|}{|EnrichedTerms_i|}$$

où $|EnrichedTerms_i|$ représente la somme de tous les termes-GO identifiés à partir des gènes

inclus dans le *cluster* i et $|SpecificEnrichedTerms_i|$ représente le nombre de termes-GO spécifiques identifiés à partir de ce même ensemble de gènes.

En se basant sur cette métrique, nous pouvons considérer une bonne classification comme produisant des *clusters* de grande taille et à fort enrichissement spécifique. Ainsi, nous avons estimé pour chaque algorithme de classification c la qualité de celle-ci (*Clustering Quality* : CQ) via la formule :

$$CQ_c = \sum_{i=1}^n SE_i * N_i$$

où N_i représente le nombre de gènes dans le *cluster* i .

Nous avons calculé le CQ pour 5 classifications : 2 basées sur ClusterONE (CO_1 et CO_2), 2 basées sur l'algorithme *fuzzy c-means* (FA_1 et FA_2) et le dernier basé sur notre modèle d'identification des *composants* par les colorations parfaites (*Component Identification* : CI). Pour CO_1 nous avons utilisé les paramètres proposés tandis que nous avons forcé l'algorithme à générer 2 *clusters* pour CO_2 . Dans le cas de FA_1 , nous avons considéré les gènes chevauchant (gènes classés dans plusieurs *clusters*) pour l'analyse d'enrichissement de gènes. Nous avons supprimé ces gènes chevauchants pour FA_2 . Enfin, pour notre méthode, en raison du fait que seuls les *composants* 2 et 6 incluait plus de 1 gène, nous avons considéré tous les autres *composants* comme non-classifiés.

TABLE 4.8 – Résultat de la comparaison avec d'autres méthodes de classification. La colonne #clusters indique sur le nombre total de *clusters*. La colonne #clusters enrichis informe sur le nombre de *clusters* enrichis. La colonne #genes indique la somme des gènes dans les *clusters* et associés à des termes-GO. Les colonnes μ^{SE} et σ^{SE} indiquent la moyenne et l'écart-type pour le SE calculés pour les *clusters* enrichis. La colonne *ratio de perte d'information* indique le pourcentage de gènes perdus lors de la classification.

Méthode de classification	#clusters	#clusters enrichis	#genes	μ^{SE}	σ^{SE}	CQ	ratio de perte d'information
CO_1	105	24	344	0.10	0.155	37.82	35.8%
CO_2	2	2	101	0.069	< 0.001	6.96	88.2%
FA_1	2	2	688	0.065	0.02	44.94	23.8%
FA_2	2	2	380	0.089	0.008	33.66	42.9%
CI	15	2	511	0.089	0.006	46.17	21.6%
Graphe PID-NCI	1	1	524	0.11	0	58.93	0%

À partir de cette comparaison (Tableau 4.8), nous pouvons noter que notre méthode d'identification des *composants* semble plus efficace pour identifier des grands *clusters* spécifiques ($CQ = 46.17$). L'enrichissement spécifique (SE) semble faible (< 0.08) dans les *clusters* obtenus. Par comparaison, lorsque nous considérons le graphe entier, le SE est de 0.11. La colonne *ratio de perte d'information* du tableau 4.8 montre une autre comparaison avec le graphe entier pour identifier les termes-GO spécifiques. Nous avons calculé cette métrique pour chaque méthode de classification c par $1 - CQ_c / CQ_{PID}$ afin d'identifier le pourcentage de gènes per-

dus lors de la classification. Ceci permet de montrer que notre méthode est celle qui obtient un meilleur score de qualité tout en ayant un ratio de perte très faible en comparaison des 4 autres méthodes de classification. De plus, la méthode de classification ayant le score le plus proche (FA_1) utilise les gènes chevauchants, ce qui amène à une hausse du CQ en raison du fait qu'un gène associé à 2 *clusters* est comptabilisé 2 fois.

4.6 Améliorations mises en place

Cette section présente 2 améliorations proposées pour la méthode des colorations parfaites. Celles-ci ont déjà été implémentées mais n'ont pas fait l'objet d'une publication contrairement aux sections précédentes.

4.6.1 Identification des nœuds corrélés en ASP

Actuellement, la méthode proposée (figure 4.3) permet d'identifier les *composants* en 3 étapes : récupération de l'intégralité des colorations parfaites. Construire pour chaque couple de nœuds, d'une matrice de corrélation afin d'identifier les nœuds corrélés et enfin fusion de ces nœuds corrélés en *composants*. L'énumération de toutes les colorations parfaites est une phase de l'analyse qui peut être particulièrement longue. De la même manière, la génération des matrices de corrélation est aussi une étape coûteuse. Nous proposons ici une amélioration de la méthode permettant de nous passer de l'énumération de ces colorations parfaites, ainsi que l'étape de la construction des matrices de corrélation. Nous utilisons pour cela une propriété de l'ASP, la recherche de l'intersection des solutions. Ainsi, il est possible de chercher, non plus un ensemble de solutions, mais une solution unique représentant l'intersection de toutes les solutions satisfaisant les contraintes définies dans le programme ASP. Nous avons donc utilisé cette propriété pour fusionner en une seule étape la recherche des colorations parfaites ainsi que l'identification des nœuds corrélés.

Pour cela, nous avons ajouté 2 règles. La première (ligne 29) exprime que 2 nœuds seront corrélés positivement s'ils ont le même signe dans une coloration. La seconde (ligne 30) exprime que 2 nœuds seront corrélés négativement s'ils ont deux signes différents dans une coloration. Notons que nous utilisons la contrainte $edge(A,B,_,_)$ afin de réduire le nombre de corrélations à identifier, réduisant ainsi la taille des solutions.

29 $correlePositif(A,B) \leftarrow coloring(A,S), coloring(B,S), edge(A,B,_,_)$.

30 $correleNegatif(A,B) \leftarrow coloring(A,S1), coloring(B,S2), S1 \neq S2, edge(A,B,_,_)$.

En cherchant l'intersection sur toutes les colorations parfaites sur ces 2 contraintes, il est alors possible d'obtenir uniquement les corrélations positives et négatives entre les nœuds. Ainsi, non seulement, on économise l'étape de génération des matrices de corrélation, mais en plus, l'espace des solutions se retrouve bien plus facile à analyser. Cette amélioration semble permettre une amélioration du temps de calcul en comparaison de l'approche précédente énumérant les colorations.

Ainsi, sur le graphe utilisé précédemment sans réduction, le temps de calcul (figure 4.3) de l'étape de l'identification des colorations parfaites à l'identification des *composants* est de 58

minutes et 35 secondes avec cette recherche de corrélations contre 114 minutes et 29 secondes en énumérant les solutions, soit un gain de près de la moitié du temps nécessaire à cette identification (Tableau 4.9, page 83). Sur ce même graphe après réduction, le gain semble équivalent. Néanmoins, l'ajout des corrélations agrandit lourdement l'espace mémoire utilisé, chaque solution contenant ainsi les colorations mais aussi les corrélations.

4.6.2 Réduction de l'espace mémoire

Nous avons ensuite cherché à réduire cette utilisation de l'espace mémoire. Du fait qu'une coloration d'un graphe de n nœuds sera composé de n tuples (nœud,signe), cela pouvait être limitant sur des analyses de grands réseaux quand le nombre de colorations est très important. En outre, il existe une redondance d'information dans le modèle présenté. En effet, la définition du graphe nous permet de connaître tous les nœuds de celui-ci. Nous associons ensuite tous ces nœuds à 2 signes possibles. Sachant que le modèle est basé sur un système à deux signes et que l'on connaît préalablement tous les nœuds, il n'est donc nécessaire que de connaître les nœuds associés à un des signes pour en déduire ceux associés à l'autre signe.

```

31 node (a, b, c) .
32 coloring (a, up) .
33 coloring (b, up) .
34 coloring (c, down) .

```

Ainsi l'exemple des lignes 31 à 34 peut être vu comme contenant la même information que celui des lignes 35 à 37. Dans le second cas, les nœuds non colorés "up" sont, par déduction, colorés "down".

```

35 node (a, b, c) .
36 coloring (a, up) .
37 coloring (b, up) .

```

Il est donc possible de reformuler la génération des solutions ainsi que les règles et contraintes de manière à travailler, non plus sur une association des signes à chaque nœud, mais la génération de toutes les combinaisons de tailles 0 à n des nœuds qui seront colorés "up". Ce formalisme nécessite donc de modifier la génération des solutions candidates (ligne 38).

```

38 0{coloringUp(I)}1 ← node(I) .

```

De plus, la définition des contraintes doit prendre en compte la possibilité qu'un nœud ne fasse pas partie de la solution candidate, que ce soit pour la définition de la coloration cohérente (lignes 39 à 42), celle de la coloration imparfaite (lignes 44 à 44) ou celle d'un régulateur pondéré imparfait (lignes 49 à 52).

```

39 consistentTarget (X) ← node (X) , coloringUp (X) , coloringUp (Y) , edge (Y, X, 1, _) .
40 consistentTarget (X) ← node (X) , coloringUp (X) , not coloringUp (Y) , edge (Y, X, -1, _) .
41 consistentTarget (X) ← node (X) , not coloringUp (X) , not coloringUp (Y) , edge (Y, X, 1, _) .
42 consistentTarget (X) ← node (X) , not coloringUp (X) , coloringUp (Y) , edge (Y, X, -1, _) .
43
44 imperfectcoloring (X) ← coloringUp (X) , not coloringUp (Z) , edge (Z, X, 1, _) .
45 imperfectcoloring (X) ← coloringUp (X) , coloringUp (Z) , edge (Z, X, -1, _) .
46 imperfectcoloring (X) ← not coloringUp (X) , coloringUp (Z) , edge (Z, X, 1, _) .
47 imperfectcoloring (X) ← not coloringUp (X) , not coloringUp (Z) , edge (Z, X, -1, _) .

```

48

```

49 imperfectWeightedRegulator(X, Y, 1..W) ← edge(X,Y,1,W), coloringUp(X), not coloringUp(Y).
50 imperfectWeightedRegulator(X, Y, 1..W) ← edge(X,Y,-1,W), coloringUp(X), coloringUp(Y).
51 imperfectWeightedRegulator(X, Y, 1..W) ← edge(X,Y,1,W), not coloringUp(X), coloringUp(Y).
52 imperfectWeightedRegulator(X, Y, 1..W) ← edge(X,Y,-1,W), not coloringUp(X), not coloringUp(Y).

```

De la même manière, la définition des corrélations entre les nœuds aussi a du être redéfinie (lignes 53 à 56).

```

53 correlePositif(X,Y) ← coloringUp(X), coloringUp(Y), edge(X,Y,_,_).
54 correlePositif(X,Y) ← not coloringUp(X), not coloringUp(Y), node(Y), node(X), edge(X,Y,_,_).
55 correleNegatif(X,Y) ← not coloringUp(X), coloringUp(Y), node(X), edge(X,Y,_,_).
56 correleNegatif(X,Y) ← coloringUp(X), not coloringUp(Y), node(Y), edge(X,Y,_,_).

```

Enfin, la définition des contraintes aussi a été modifiée afin de prendre en compte cette nouvelle génération de solutions candidates.(lignes 57 à 67).

```

57 inconsistentTarget(X) ← not consistentTarget(X), target(X).
58 sumInconsistencyTargets(X) ← X=#count{ node(Z) :inconsistentTarget(Z) }.
59 #minimize {X@3 : sumInconsistencyTargets(X)}.
60
61 sumImperfectcoloring(X) ← X=#count{ node(Z) :imperfectcoloring(Z) }.
62 #minimize {X@2 : sumImperfectcoloring(X)}.
63
64 sumImperfectWeightedRegulatorPerTarget(X,Y) ← Y=#count{ x(A, B) :
    imperfectWeightedRegulator(A,X,B)}, imperfectcoloring(X).
65 imperfectWeightedRegulatorPerTarget(X,1..W) ← sumImperfectWeightedRegulatorPerTarget(X,W).
66 sumImperfectWeightedRegulator(X) ← X=#count{ x(Y,Z) :
    imperfectWeightedRegulatorPerTarget(Y,Z),imperfectcoloring(Y)}.
67 #minimize {X@1 : sumImperfectWeightedRegulator(X)}.

```

Nous avons initialement supposé que cette méthode de génération des solutions candidates aurait permis de réduire le temps de calcul en limitant les opérations d'affectation. Néanmoins, il est à noter que dans ce nouveau formalisme, le temps de calcul semble légèrement plus important que l'approche précédente à 2 signes sur le graphe compacté (Tableau 4.9), soit 21 secondes contre 14. Ceci est probablement dû à l'ajout de contraintes afin de prendre en compte ce nouveau formalisme et qui sont à vérifier à chaque étape. Néanmoins, le gain de temps devient positif dès que l'on travaille sur ce même graphe non-compacté, permettant de réduire le temps de calcul de 58 minutes et 35 secondes à un temps de 48 minutes et 15 secondes. De plus, l'espace mémoire utilisé semble être réduit, permettant d'analyser localement certains réseaux plus importants.

4.6.3 Amélioration de la réduction du graphe

Nous avons présenté précédemment les motifs recherchés dans le graphe afin de le réduire en identifiant des nœuds qui appartiendraient au même *composant*. Dans le modèle présenté, nous

TABLE 4.9 – Temps de calcul des 3 méthodes d’identification des *composants*. La méthode d’énumération des colorations est celle présentée en section 4.2.1 et qui a servi à l’application. La méthode de corrélation des colorations est décrite à la section 4.6.1. La méthode à 1 signe est celle décrite dans la section 4.6.2.

Méthode	Graphe	#Nœuds	#Arcs	Nombre de <i>composants identifiés</i>	Temps de calcul
Énumération des colorations	Origine	2269	2683	15	114 min 29 sec
Corrélation des colorations	Origine	2269	2683	15	58 min 35 sec
Modèle à 1 signe	Origine	2269	2683	15	48 min 15 sec
Énumération des colorations	Compacté	193	389	15	35 sec
Corrélation des colorations	Compacté	193	389	15	14 sec
Modèle à 1 signe	Compacté	193	389	15	21 sec

cherchions successivement ces motifs mais cette recherche n’était effectuée qu’une seule fois. Or, la fusion de nœuds peut amener à l’apparition d’un nouveau motif qui, potentiellement, ne sera pas détecté. C’est dans ce but que nous avons modifié l’algorithme d’identification de ces motifs afin de les rechercher tant que des motifs étaient trouvés.

Une seconde amélioration a porté sur un des motifs recherchés. En effet, dans le cas de la réduction basée sur les co-régulateurs (figure 4.5-b), il est possible aussi de considérer qu’un nœud n n’ayant qu’un successeur s et un seul arc avec celui-ci sera corrélé avec ce successeur. De plus, la coloration de s sera cohérente dans les colorations parfaites car n suffira à l’expliquer. Il est alors possible de fusionner n et s en stockant l’information que la coloration de s est cohérente.

Notons néanmoins que dans le cas du graphe que nous avons utilisé dans l’application, cette amélioration n’a pas permis de réduire plus fortement la taille du graphe. En effet, la génération du graphe à partir de 3 nœuds sources par les chemins les plus courts vers les gènes variants n’a pas amené à ce type de motifs. Néanmoins, nous avons pu appliquer ce modèle de réduction amélioré sur la deuxième version de la base de données Trrust (<http://www.grnpedia.org/trrust/>) qui référence 2048 gènes humains connectés par 5060 arcs (activateurs et inhibiteurs). Si la première méthode de réduction permettait de réduire le graphe initial à une taille de 1192 nœuds et 3699 arcs, la nouvelle version de réduction a permis de le réduire à une taille de 1053 nœuds pour 3602 arcs.

4.7 Conclusion

Dans ce chapitre, nous avons proposé une méthode qui impose des contraintes pour modéliser la coloration de graphes sur un réseau de régulation biologique. Cette méthode permet de réduire un graphe en sous-graphes appelés *composants*. Ces *composants* décrivent les colorations du réseau indépendantes les unes des autres sous les contraintes de colorations parfaites. De plus, en intégrant des observations, nous pouvons sélectionner certains de ces *composants* en se basant sur la similarité maximale entre les configurations de colorations des *composants* et ces observations. Contrairement aux autres méthodes d’extraction de sous-graphes, la méthode présentée ici les identifie en se basant uniquement sur la logique du réseau (causalité et

interactions activatrices/inhibitrices). En outre, la méthode extrait les sous-graphes en se basant sur les colorations parfaites, puis seulement intègre les données d'expression en les confrontant aux configurations de coloration. Cette particularité pourrait limiter les biais liés à la spécificité des données. Enfin, nous intégrons ici les données d'expression dans le réseau au niveau transcriptionnel. L'intégration de ces données d'observations permet d'évaluer la similarité entre un profil d'expression et les configurations parfaites de chaque *composant*. Ainsi, nous pouvons identifier des profils ou groupes de profils dont les gènes semblent avoir des expressions très éloignées de ces configurations parfaites.

Au vu des résultats obtenus lors de l'application sur données réelles, cette méthode semble efficace pour identifier et sélectionner des *composants* spécifiques à des profils d'expression. Nous avons ainsi pu identifier un *composant* sur lequel les observations d'expression de gènes des profils NCP et MC étaient significativement différentes. De plus, au vu de la similarité calculée, les profils MC avaient une similarité plus faible que les NPC, semblant indiquer que ce *composant* était plus dérégulé chez ces profils. Ces *composants* pouvant être liés à des fonctions biologiques, nous avons pu établir un lien entre des fonctions oncogéniques et ce *composant* associé aux profils MC. Ce résultat semble cohérent avec la nature des données analysées. Enfin, nous avons comparé notre méthode avec d'autres approches identifiant des sous-graphes. Pour cela, nous avons introduit une métrique permettant d'évaluer la qualité d'un *clustering* de graphe en se basant sur les GO-termes et leur niveau. A partir de celle-ci, nous avons pu montrer que notre méthode des colorations parfaites était celle qui générait des *clusters* de meilleure qualité en comparaison des autres approches utilisées.

Bien que cette méthode semble efficace sur des réseaux de grande taille, elle reste néanmoins limitée sur des modèles plus importants. Ainsi, travailler sur la base NCI-PID en entier n'a pas été possible (temps de calcul supérieur à une semaine). Dans cette optique, nous avons mis au point plusieurs améliorations de cette méthode afin de réduire le temps de calcul et l'espace mémoire utilisé. Nous présenterons en fin de manuscrit une possibilité d'amélioration de cette méthode afin d'intégrer des données non plus discrètes mais continues. Cette évolution permettrait de maintenir la possibilité d'analyser de grands réseaux de régulation biologiques par une modélisation à 2 signes de ceux-ci tout en évitant la perte d'information consécutive à une étape de discrétisation.

Conclusion & perspectives

5.1 Contributions

Cette section reprend les diverses contributions présentées dans ce manuscrit. Une liste plus détaillée des productions et communications scientifiques est trouvable à l'annexe .1.

5.1.1 La coloration cohérente des graphes

Le modèle de coloration cohérente des graphes permet, à partir d'un réseau de régulation et un profil d'expression contenant des colorations de nœuds, d'inférer les colorations des autres nœuds. Cette inférence se base sur des règles logiques assurant que chaque coloration soit justifiée par, au-moins, un prédécesseur. Dans le cas où il n'est pas possible de justifier toutes les colorations, ce modèle permet d'identifier les réparations à effectuer, soit sur le graphe (MCOS-repairs) ou sur les données (SCENFIT).

Nous avons appliqué ce modèle de coloration des graphes sur des données transcriptomiques issues du myélome multiple et avons proposé une méthode de comparaison des colorations inférées afin de comparer plusieurs profils entre eux. Nous avons travaillé avec les données de 611 jeux de données dont 602 issus de cellules de myélome multiple et 9 de issus de cellules saines avec un réseau de régulation extrait de la base de données NCI-PID. La confrontation entre ces données et le réseaux de régulation, avec une correction MCOS-repairs (correction de graphe) a permis d'identifier l'état des éléments non-observés de ce réseau. L'originalité de ce travail repose sur l'utilisation de ces données inférées pour chercher les colorations les plus discriminantes entre les échantillons malades et sains. Les 2 approches utilisées, comparaison de fréquences et arbres de décision, nous ont permis d'identifier à chaque fois les couples (JUN/FOS[n],-) et (FOXMI*[c],-) comme associés à la majorité des cellules cancéreuses. De plus, il a été possible de lier ces colorations avec les connaissances actuelles sur les dérégulations des voies de régulation dans le myélome multiple, permettant de valider la cohérence de ces prédictions. Enfin, en se basant sur ces colorations inférées, nous avons pu identifier

des marqueurs de survie et enrichir un modèle pronostique basé sur les données cliniques des patients.

Ensuite, nous avons proposé un modèle issu de la coloration cohérente de graphes, permettant, via une correction SCENFIT-repairs (correction des observations) de simuler l'impact d'une perturbation sur un réseau de régulation et un profil d'expression. Ce modèle reste limité car ne pouvant simuler qu'une perturbation à la fois. Ce chapitre a fait l'objet d'une présentation [12] et de plusieurs posters [11, 13, 14] et a été accepté dans le journal Scientific-Reports [93].

5.1.2 Les colorations parfaites

A la suite de ces travaux, nous avons proposé un nouveau modèle de recherche de colorations parfaites. Celui-ci permet, à partir d'un réseau de régulation, d'identifier les ensembles de nœuds de ce réseau dont les colorations seront corrélées sous les contraintes de ce modèle. Ces nœuds corrélés, positivement ou négativement, sont réunis au sein d'un même *composant*. Cette méthode permet aussi d'intégrer des données d'expression de gènes et de calculer la similarité maximale (MS) entre les observations des nœuds d'un *composant* et les configurations parfaites de celui-ci.

L'application de ce modèle sur les mêmes données que pour le modèle de colorations cohérentes nous a permis d'identifier 15 *composants* différents dont 2 contenant plus de 2 gènes. Sur ces 2 *composants*, le MS du premier (C_2) a pu être identifié comme significativement différent entre les cellules normales et cancéreuses, amenant à supposer une dérégulation forte du réseau formant ce *composant*. De plus celui-ci a pu être associé à des fonctions oncogéniques. Enfin, afin de valider notre méthode, nous nous sommes comparés avec d'autres approches de classification de graphes. Cette comparaison a permis de mettre en évidence qu'utiliser les colorations parfaites pour identifier des sous-graphes semblait plus apte à regrouper des gènes partageant des fonctions biologiques spécifiques, renforçant ainsi nos résultats initiaux. Ces résultats ont été présentés au *workshop* CNB-MAC, attendant à la conférence ACM-BCM 2017 publié sous forme de *proceedings* [94]. Ils sont actuellement en cours de relecture pour le journal BMC Systems Biology.

Enfin, afin de permettre d'analyser plus efficacement de grands réseaux, nous avons présenté 3 améliorations successives de la méthode. La première est basée sur un rassemblement des étapes de recherche des nœuds corrélés et de la recherche des colorations parfaites en une seule phase. La seconde se base sur une optimisation de l'espace mémoire utilisé en réduisant la redondance d'informations des colorations. Enfin, la dernière amélioration permet d'identifier de nouveaux motifs dans le graphe et ainsi de réduire plus efficacement celui-ci avant la recherche des nœuds corrélés. Ces 3 améliorations sont cumulables, amenant à des gains de plus de la moitié du temps de calcul.

5.2 Perspectives

Nous présentons ici deux possibilités de poursuite de nos travaux. Chacun de ces axes est la continuité d'une des contributions présentées précédemment. Une troisième possibilité qui ne sera pas développée plus longuement ici serait l'application des méthodes présentées et amenées

à être améliorée, dans des contextes plus ouverts afin de les faire connaître. Ainsi, récemment, le *Multiple Myeloma Dream Challenge* a été proposé, présentant plusieurs sous-compétitions, dont une sur de la prédiction de la progression de la maladie à 18 mois, basée sur des données d'expression de gènes. Ce type de projets, de par les données initiales et les objectifs recherchés, pourrait être tout à fait adapté pour une application des méthodes présentées précédemment.

5.2.1 Un modèle de perturbations multiples

En se basant sur la coloration cohérente des graphes, nous avons exploré la possibilité de simuler l'impact d'une perturbation sur un réseau de régulation et un profil d'expression. Cette méthode restait limitée car ne simulant qu'une perturbation à la fois, et n'évaluant l'impact de celle-ci que sur le profil d'expression. Cela étant, elle ouvrait la possibilité d'introduire ce type de simulation dans des approches d'identification de cibles thérapeutiques, ou de compréhension des mécanismes de régulation cellulaires [61]. Nous pouvons ici imaginer un modèle (figure 5.1) qui se baserait sur les colorations inférées à partir d'un réseau de régulation et d'un profil d'expression. Ce modèle prendrait en entrée ces colorations, une liste de perturbations possibles et un motif à atteindre afin d'identifier la ou les combinaisons de perturbations à effectuer pour obtenir le motif (ensemble de colorations) souhaité.

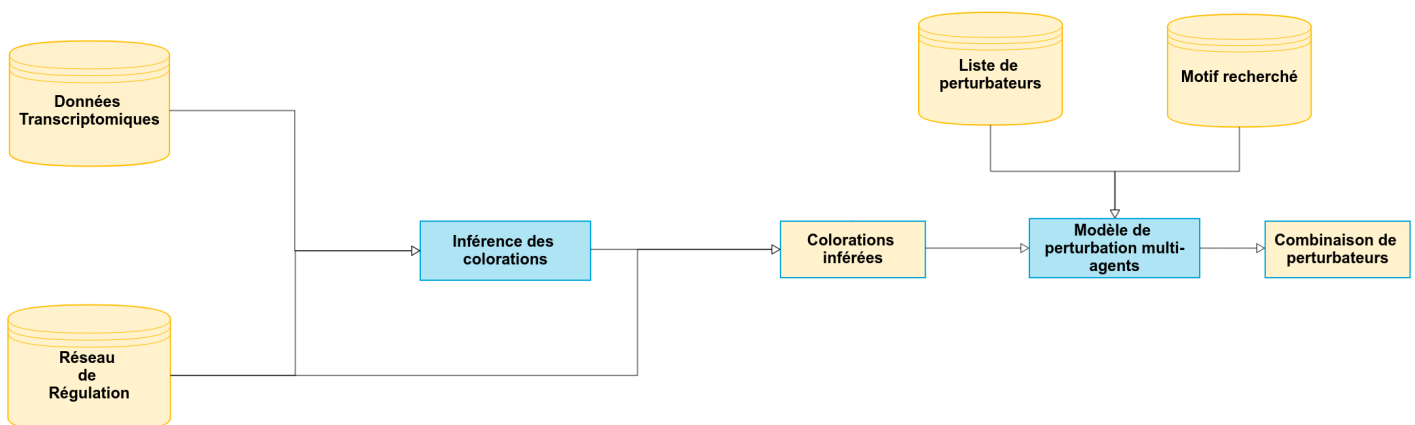


FIGURE 5.1 – Modèle de perturbations multiples.

Pour cela, il pourrait être intéressant de chercher pour chaque combinaison de perturbateurs son impact non plus sur le profil d'expression mais sur l'intégralité des colorations prédites. Il serait ensuite possible d'estimer si cette combinaison amène à une nouvelle coloration correspondant à un motif pré-déterminé.

Un modèle s'en approchant (voir annexe .3) a déjà été développé permettant à partir d'un graphe, de sa coloration, d'une liste de perturbations possibles et d'un motif à obtenir, d'identifier la combinaison minimale de perturbations à appliquer pour amener à ce motif pré-déterminé (Exemple d'application à la figure 5.2).

Il est à noter que ce modèle n'a pas été testé avec des données réelles. De plus, celui-ci ne prend en compte qu'une coloration et n'intègre pas les réparations potentielles du graphe lors de l'inférence des colorations. Néanmoins, il pourrait permettre d'améliorer la modélisation

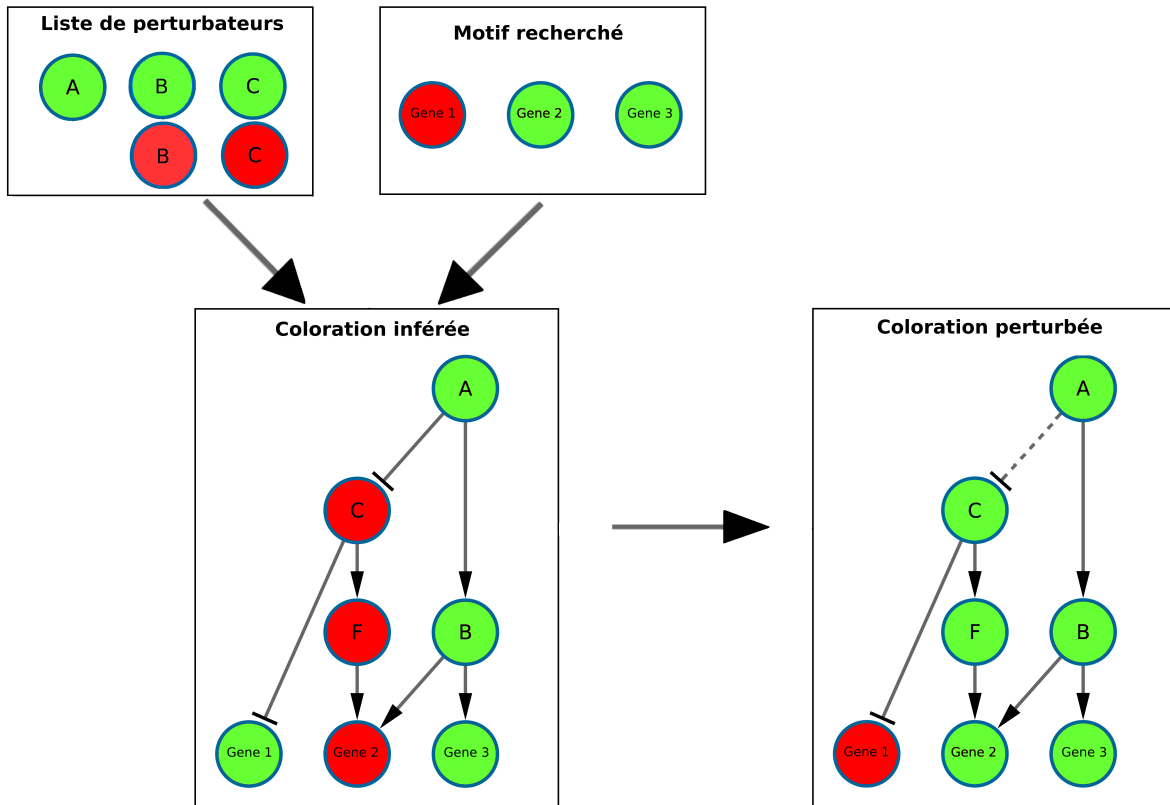


FIGURE 5.2 – Exemple d’application du modèle de perturbations multiples. Dans cet exemple, le modèle identifiera la perturbation "(c,-)" comme suffisante pour amener au motif recherché.

des perturbations et donc de répondre plus efficacement à la recherche des cibles thérapeutiques personnalisées et ainsi ouvrir une perspective plus appliquée des méthodes de coloration.

5.2.2 L’intégration de données continues dans les colorations parfaites

Il a été montré en début de ce manuscrit la difficulté que peut poser la phase de discrétisation des données. En effet, de nombreuses méthodes existent (quelques uns ont été présentées dans le paragraphe 1.5.2), chacune présentant des avantages et inconvénients, eux-mêmes dépendant du contexte de leur utilisation. Néanmoins, cette phase est rendu nécessaire pour étudier de grandes quantités de données malgré la perte d’informations obligatoirement associée.

Dans les études présentées, nous avons utilisé une discrétisation par seuil, calculée afin de maximiser à posteriori la précision de la méthode de coloration cohérente des graphes. Dans le cas de ce modèle de colorations cohérentes des graphes, nous avons essayé de contrebalancer en partie la perte d’information de la discrétisation en intégrant la non-variation comme signe possible pour un gène.

Néanmoins, dans la méthode des colorations parfaites que nous avons présentée, ces colorations sont basées sur un modèle à 2 signes pour identifier les *composants*. De même, si celle-ci permet d’intégrer des données d’expression de gènes afin d’estimer le niveau de dérégulation des *composants* vis à vis des colorations parfaites, les gènes non-variants n’ont ici pas d’impact sur cette métrique et les observations de gènes sont, là aussi, considérées via 2

signes possibles. Ceci conduit de fait à une réduction des capacités à appréhender finement les phénomènes de régulation dans les réseaux biologiques. De même, si la non-variation est une information, une faible variation peut, elle aussi, apporter une information potentiellement intéressante. Aussi, nous pensons qu'un axe d'amélioration important de ce modèle de colorations parfaites pourrait être l'intégration de données continues dans celui-ci. En effet, si la recherche des colorations parfaites se base sur un modèle à 2 signes, il est tout à fait envisageable d'intégrer des données en passant non plus par une étape de discrétisation mais de normalisation sur un intervalle donné dont la borne inférieure (respectivement supérieure) correspondrait à l'état le plus inhibé (respectivement activé) des expressions de gènes. Il serait alors possible de considérer une modification du modèle présenté dans le chapitre 4 (figure 5.3)

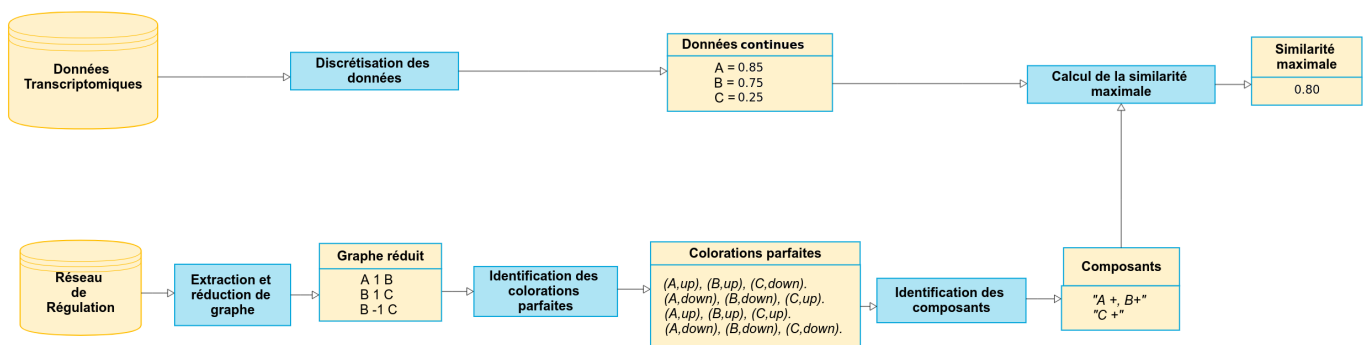


FIGURE 5.3 – Version modifiée du *workflow* de la figure 4.3 pour intégrer des données continues.

Il serait alors possible de caractériser un profil d'expression non plus sur la base des gènes identifiés comme sur et sous-activés, mais sur l'intégralité des mesures de celui-ci. Cette autre forme d'intégration permettrait ainsi de profiter des avantages d'un modèle statique discret dans la recherche des colorations parfaites, en particulier au niveau du temps de calcul et du coût mémoire tout en réduisant la perte d'information liée à la discrétisation à une étape de normalisation. Néanmoins, cette normalisation occasionnera, là aussi, une homogénéisation artificielle des données, amenant là aussi à un axe de recherche à explorer. Toutefois, cette étape ne changerait pas fondamentalement le modèle initial, car elle permettrait de comparer les apports des 2 types de données (continues et discrètes) tout en profitant des méthodes de réduction de graphes précédemment présentées et déjà implémentées.

5.2.3 Perspectives générales

De manière plus globale, les méthodes présentées utilisaient des données d'expression de gènes intégrées à des réseaux de régulation. Bien que le fait de travailler sur les ARN et non l'ADN des cellules cancéreuses permette de limiter le bruit lié aux mutations "spectatrices" [114], il existe encore un décalage avec le protéome lié aux phénomènes post-traductionnels. Néanmoins, il n'est pas possible à l'heure actuelle d'avoir des données sur les protéines par manque de matériel tumoral et ce travail présentait l'avantage d'inférer ce type d'informations à partir de données ARN et des réseaux de régulation. Il pourrait, toutefois, être intéressant d'intégrer ce type de

phénomène, dans ces modèles. De la même manière, les modèles présentés considèrent une population homogène de cellules cancéreuses indépendantes de leur environnement. Ainsi, nous n'intégrons pas la diversité des cellules cancéreuses au sein d'une même tumeur [87], ni les interactions entre ces cellules et celles de leur milieu. Ces derniers axes pourraient, là aussi, servir de base à des travaux futurs de recherche.

Bibliographie

- [1] E. B. Abdallah, T. Ribeiro, M. Magnin, O. Roux, and K. Inoue. *Reducing the search space in literature-based discovery by exploring outlier documents a case study in finding links between gut microbiome and Alzheimer's disease*, volume 3. jan 2017. [28](#)
- [2] C. Ainali, N. Valeyev, G. Perera, A. Williams, J. E. Gudjonsson, C. a. Ouzounis, F. O. Nestle, and S. Tsoka. Transcriptome classification reveals molecular subtypes in psoriasis. *BMC genomics*, 2012. [21](#)
- [3] B. L. Aken, S. Ayling, D. Barrell, L. Clarke, V. Curwen, S. Fairley, J. Fernandez Banet, K. Billis, C. García Girón, T. Hourlier, K. Howe, A. Kähäri, F. Kokocinski, F. J. Martin, D. N. Murphy, R. Nag, M. Ruffier, M. Schuster, Y. A. Tang, J.-H. Vogel, S. White, A. Zaidissa, P. Flicek, and S. M. J. Searle. The Ensembl gene annotation system. *Database : the journal of biological databases and curation*, 2016, 2016. [13](#)
- [4] C. M. Annunziata, R. E. Davis, Y. Demchenko, W. Bellamy, A. Gabrea, F. Zhan, G. Lenz, I. Hanamura, G. Wright, W. Xiao, S. Dave, E. M. Hurt, B. Tan, H. Zhao, O. Stephens, M. Santra, D. R. Williams, L. Dang, B. Barlogie, J. D. Shaughnessy, W. M. Kuehl, and L. M. Staudt. Article Frequent Engagement of the Classical and Alternative NF- k B Pathways by Diverse Genetic Abnormalities in Multiple Myeloma. (August) :115–130, 2007. [32](#)
- [5] H. Avet-Loiseau, M. Attal, P. Moreau, C. Charbonnel, F. Garban, C. Hulin, S. Leyvraz, M. Michallet, I. Yakoub-Agha, L. Garderet, G. Marit, L. Michaux, L. Voillat, M. Renaud, B. Grosbois, G. Guillermin, L. Benboubker, M. Monconduit, C. Thieblemont, P. Casasus, D. Caillot, A.-M. Stoppa, J.-J. Sotto, M. Wetterwald, C. Dumontet, J.-G. Fuzibet, I. Azais, V. Dorvaux, M. Zandecki, R. Bataille, S. Minvielle, J.-L. Harousseau, T. Facon, and C. Mathiot. Genetic abnormalities and survival in multiple myeloma : the experience of the Intergroupe Francophone du Myélome. *Blood*, 109(8) :3489–95, apr 2007. [32](#)
- [6] C. Backes, A. Keller, J. Kuentzer, B. Kneissl, N. Comtesse, Y. A. Elnakady, R. Müller, E. Meese, and H.-P. Lenhof. GeneTrail–advanced gene set enrichment analysis. *Nucleic acids research*, 35(Web Server issue) :W186–92, jul 2007. [35](#), [36](#)
- [7] C. Backes, A. Rurainski, G. W. Klau, O. Müller, D. Stöckel, A. Gerasch, J. Kuntzer, D. Maisel, N. Ludwig, M. Hein, A. Keller, H. Burtscher, M. Kaufmann, E. Meese, and H.-P. Lenhof. An integer linear programming approach for finding deregulated subgraphs in regulatory networks. *Nucleic acids research*, 40(6) :e43, mar 2012. [36](#), [37](#), [63](#)

- [8] C. Baral. *Knowledge Representation, Reasoning and Declarative Problem Solving*. Cambridge University Press, 2003. [23](#)
- [9] A. D. Baxevanis. The Molecular Biology Database Collection : an updated compilation of biological database resources. *Nucleic acids research*, 29(1) :1–10, jan 2001. [12](#)
- [10] G. A. Bazykin and A. V. Kochetov. Alternative translation start sites are conserved in eukaryotic genomes. *Nucleic acids research*, 39(2) :567–77, jan 2011. [9](#)
- [11] Bertrand Miannay, Florence Magrangeas, Olivier Roux, Stephane Minvielle, and Carito Guziolowski. Key transcription factors altered in multiple myeloma patients revealed by logic programming approach combining gene expression profiling and regulatory networks. EMBO conference : From Functional Genomics to Systems Biology, Heidelberg, Germany, November 12–15, 2016. [86](#), [106](#)
- [12] Bertrand Miannay, Olivier Roux, Carito Guziolowski, Stéphane Minvielle, Florence Magrangeas . Identification des voies de signalisation impliquées dans le myélome multiple par programmation par contrainte. Bioss seminary, Lyon, France, July 1–2, 2016. [86](#)
- [13] Bertrand Miannay, Stephane Minvielle, Morgan Magnin, Florence Magrangeas, and Carito Guziolowski . Understanding myelome multiple mechanisms by automatic reasoning on an integrated model of transcriptomic data and large-scale signaling pathways. CMSB-2015 Conference, Nantes France, September 16–18, 2015. [86](#), [106](#)
- [14] Bertrand Miannay, Stephane Minvielle, Morgan Magnin, Florence Magrangeas, and Carito Guziolowski . Understanding myelome multiple patients relapse by automatic reasoning on an integrated model of transcriptomic data and large-scale signaling pathways. RECOMB/ISCB Conference on Regulatory and Systems Genomics, San Diego, United States of America, November 9–14, 2014. [86](#), [105](#)
- [15] G. Bindea, B. Mlecnik, H. Hackl, P. Charoentong, M. Tosolini, A. Kirilovsky, W.-H. Fridman, F. Pagès, Z. Trajanoski, and J. Galon. ClueGO : a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics (Oxford, England)*, 25(8) :1091–3, may 2009. [33](#)
- [16] B. Boeckmann, A. Bairoch, R. Apweiler, M.-C. Blatter, A. Estreicher, E. Gasteiger, M. J. Martin, K. Michoud, C. O’Donovan, I. Phan, S. Pilbout, and M. Schneider. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic acids research*, 31(1) :365–70, jan 2003. [14](#)
- [17] R. J. Bold, P. M. Termuhlen, and D. J. McConkey. Apoptosis, cancer and cancer therapy. *Surgical Oncology*, 6(3) :133–142, nov 1997. [76](#)
- [18] N. Bolli, H. Avet-Loiseau, D. C. Wedge, P. Van Loo, L. B. Alexandrov, I. Martincorena, K. J. Dawson, F. Iorio, S. Nik-Zainal, G. R. Bignell, J. W. Hinton, Y. Li, J. M. C. Tubio, S. McLaren, S. O’ Meara, A. P. Butler, J. W. Teague, L. Mudie, E. Anderson, N. Rashid,

- Y.-T. Tai, M. a. Shamma, A. S. Sperling, M. Fulciniti, P. G. Richardson, G. Parmigiani, F. Magrangeas, S. Minvielle, P. Moreau, M. Attal, T. Facon, P. A. Futreal, K. C. Anderson, P. J. Campbell, and N. C. Munshi. Heterogeneity of genomic evolution and mutational profiles in multiple myeloma. *Nature communications*, 5 :2997, 2014. [32](#)
- [19] S. Boué, M. Talikka, J. W. Westra, W. Hayes, A. Di Fabio, J. Park, W. K. Schlage, A. Sewer, B. Fields, S. Ansari, F. Martin, E. Veljkovic, R. Kenney, M. C. Peitsch, and J. Hoeng. Causal biological network database : a comprehensive platform of causal biological network models focused on the pulmonary and vascular systems. *Database : the journal of biological databases and curation*, 2015 :bav030, jan 2015. [13](#)
- [20] M. Carlebach, A. Amiel, E. Gaber, J. Radnay, Y. Manor, M. Fejgin, and M. Lishner. Multiple myeloma : monoallelic deletions of the tumor suppressor genes tp53 and rb1 in long-term follow-up. *Cancer genetics and cytogenetics*, 117(1) :57–60, 2000. [30](#)
- [21] A. Cassese, M. Guindani, P. Antczak, F. Falciani, and M. Vannucci. A Bayesian model for the identification of differentially expressed genes in *Daphnia magna* exposed to munition pollutants. *Biometrics*, 71(3) :803–11, sep 2015. [28](#)
- [22] J. Catlett. On changing continuous attributes into ordered discrete attributes. In *Machine Learning — EWSL-91*, pages 164–178. Springer-Verlag, Berlin/Heidelberg, 1991. [28](#)
- [23] E. G. Cerami, B. E. Gross, E. Demir, I. Rodchenkov, O. Babur, N. Anwar, N. Schultz, G. D. Bader, and C. Sander. Pathway Commons, a web resource for biological pathway data. *Nucleic acids research*, 39(Database issue) :D685–90, jan 2011. [15](#)
- [24] L. Chen, S. Wang, Y. Zhou, X. Wu, I. Entin, J. Epstein, S. Yaccoby, W. Xiong, B. Barlogie, J. D. Shaughnessy, and F. Zhan. Identification of early growth response protein 1 (EGR-1) as a novel target for JUN-induced apoptosis in multiple myeloma. *Blood*, 115(1) :61–70, jan 2010. [47](#), [52](#)
- [25] L. T. Chow, R. E. Gelinas, T. R. Broker, and R. J. Roberts. An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. *Cell*, 12(1) :1–8, sep 1977. [9](#)
- [26] M.-L. Chretien, J. Corre, V. Lauwers-Cances, F. Magrangeas, A. Cleyne, E. Yon, C. Hulin, X. Leleu, F. Orsini-Piocelle, J.-S. Blade, C. Sohn, L. Karlin, X. Delbrel, B. Hebraud, M. Roussel, G. Marit, L. Garderet, M. Mohty, P. Rodon, L. Voillat, B. Royer, A. Jaccard, K. Belhadj, J. Fontan, D. Caillot, A.-M. Stoppa, M. Attal, T. Facon, P. Moreau, S. Minvielle, and H. Avet-Loiseau. Understanding the role of hyperdiploidy in myeloma prognosis : which trisomies really matter ? *Blood*, 126(25), 2015. [32](#)
- [27] K. R. Christie, S. Weng, R. Balakrishnan, M. C. Costanzo, K. Dolinski, S. S. Dwight, S. R. Engel, B. Feierbach, D. G. Fisk, J. E. Hirschman, E. L. Hong, L. Issel-Tarver, R. Nash, A. Sethuraman, B. Starr, C. L. Theesfeld, R. Andrada, G. Binkley, Q. Dong,

- C. Lane, M. Schroeder, D. Botstein, and J. M. Cherry. Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from *Saccharomyces cerevisiae* and related sequences from other organisms. *Nucleic acids research*, 32(Database issue) :D311–4, jan 2004. [13](#)
- [28] W. F. W. F. Clocksin and C. S. C. S. Mellish. *Programming in Prolog*. Springer-Verlag, 2003. [23](#)
- [29] A. M. L. Coluccia, T. Cirulli, P. Neri, D. Mangieri, M. C. Colanardi, A. Gnoni, N. Di Renzo, F. Dammacco, P. Tassone, D. Ribatti, C. Gambacorti-Passerini, and A. Vacca. Validation of PDGFRbeta and c-Src tyrosine kinases as tumor/vessel targets in patients with multiple myeloma : preclinical efficacy of the novel, orally available inhibitor dasatinib. *Blood*, 112(4) :1346–56, aug 2008. [47](#)
- [30] I. C. G. Consortium et al. International network of cancer genome projects. *Nature*, 464(7291) :993, 2010. [13](#)
- [31] C. D. Davis and E. O. Uthus. DNA Methylation, Cancer Susceptibility, and Nutrient Interactions. *Experimental Biology and Medicine*, 229(10) :988–995, nov 2004. [10](#)
- [32] O. Decaux, L. Lodé, F. Magrangeas, C. Charbonnel, W. Gouraud, P. Jézéquel, M. Attal, J. L. Harousseau, P. Moreau, R. Bataille, L. Campion, H. Avet-Loiseau, and S. Minvielle. Prediction of survival in multiple myeloma based on gene expression profiles reveals cell cycle and chromosomal instability signatures in high-risk patients and hyperdiploid signatures in low-risk patients : A study of the Intergroupe Francophone du Myélome. *Journal of Clinical Oncology*, 26(29) :4798–4805, oct 2008. [32](#), [50](#)
- [33] J. DeRisi, L. Penland, P. O. Brown, M. L. Bittner, P. S. Meltzer, M. Ray, Y. Chen, Y. A. Su, and J. M. Trent. Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nature Genetics*, 14(4) :457–460, dec 1996. [11](#), [27](#)
- [34] S. Draghici, P. Khatri, A. L. Tarca, K. Amin, A. Done, C. Voichita, C. Georgescu, and R. Romero. A systems biology approach for pathway level analysis. *Genome research*, 17(10) :1537–45, Oct. 2007. [36](#), [37](#)
- [35] P. Drouin. Etude des méthodes de discrétisation pour les données d’expression génique. Technical report, 2016. [27](#)
- [36] M. Elati, P. Neuvial, M. Bolotin-Fukuhara, E. Barillot, F. Radvanyi, and C. Rouveirol. LICORN : learning cooperative regulation networks from gene expression data. *Bioinformatics*, 23(18) :2407–2414, sep 2007. [29](#), [36](#), [37](#), [63](#)
- [37] E. Elcock. Absys : the first logic programming language —A retrospective and a commentary. *The Journal of Logic Programming*, 9(1) :1–17, jul 1990. [23](#)
- [38] J. Elder. *Handbook of statistical analysis and data mining applications*. Academic Press, 2009. [21](#)

- [39] F. E. Faisal and T. Milenkovi ? Dynamic networks reveal key players in aging. *Bioinformatics*, 30(12) :1721–1729, jun 2014. [34](#), [36](#), [37](#), [63](#)
- [40] F. Fan, G. Tonon, M. H. Bashari, S. Vallet, E. Antonini, H. Goldschmidt, H. Schulze-Bergkamen, J. T. Opferman, M. Sattler, K. C. Anderson, D. Jäger, and K. Podar. Targeting Mcl-1 for multiple myeloma (MM) therapy : drug-induced generation of Mcl-1 fragment Mcl-1(128-350) triggers MM cell death via c-Jun upregulation. *Cancer letters*, 343(2) :286–94, feb 2014. [47](#)
- [41] J. Ferlay, E. Steliarova-Foucher, J. Lortet-Tieulent, S. Rosso, J. W. W. Coebergh, H. Comber, D. Forman, and F. Bray. Cancer incidence and mortality patterns in Europe : Estimates for 40 countries in 2012. *European Journal of Cancer*, 49 :1374–1403, 2013. [31](#)
- [42] A. Z. Fire. Gene silencing by double-stranded RNA. *Cell Death and Differentiation*, 14(12) :1998–2012, dec 2007. [10](#)
- [43] M. Y. Galperin and X. M. Fernandez-Suarez. The 2012 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection. *Nucleic Acids Research*, 40(D1) :D1–D8, jan 2012. [13](#)
- [44] M. Y. Galperin, X. M. Fernández-Suárez, and D. J. Rigden. The 24th annual Nucleic Acids Research database issue : a look back and upcoming changes. *Nucleic acids research*, 45(D1) :D1–D11, jan 2017. [14](#)
- [45] M. Gebser, R. Kaminski, B. Kaufmann, and T. Schaub. Answer Set Solving in Practice. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 6(3) :1–238, dec 2012. [24](#)
- [46] M. Gebser, T. Schaub, S. Thiele, B. Usadel, and P. Veber. Detecting Inconsistencies in Large Biological Networks with Answer Set Programming. pages 130–144. Springer, Berlin, Heidelberg, 2008. [36](#)
- [47] C. Goble and R. Stevens. State of the nation in data integration for bioinformatics. *Journal of Biomedical Informatics*, 41(5) :687–693, oct 2008. [13](#)
- [48] D. Gomez-Cabrero, I. Abugessaisa, D. Maier, A. Teschendorff, M. Merkenschlager, A. Gisel, E. Ballestar, E. Bongcam-Rudloff, A. Conesa, and J. Tegnér. Data integration in the era of omics : current and future challenges. *BMC systems biology*, 8(2) :11, 2014. [12](#)
- [49] K. A. Gray, B. Yates, R. L. Seal, M. W. Wright, and E. A. Bruford. Genenames.org : the HGNC resources in 2015. *Nucleic Acids Research*, 43(D1) :D1079–D1085, jan 2015. [14](#)
- [50] C. Gu, Y. Yang, R. Sompallae, H. Xu, V. S. Tompkins, C. Holman, D. Hose, H. Goldschmidt, G. Tricot, F. Zhan, and S. Janz. FOXM1 is a therapeutic target for high-risk multiple myeloma. *Leukemia*, dec 2015. [32](#), [47](#), [48](#)

- [51] G. Guo, F. von Meyenn, M. Rostovskaya, J. Clarke, S. Dietmann, D. Baker, A. Sahakyan, S. Myers, P. Bertone, W. Reik, K. Plath, and A. Smith. Epigenetic resetting of human pluripotency. *bioRxiv*, 2017. [10](#)
- [52] A. Gusnanto, S. Calza, and Y. Pawitan. Identification of differentially expressed genes and false discovery rate in microarray studies. *Current Opinion in Lipidology*, 18(2) :187–193, apr 2007. [27](#)
- [53] A. H. F. G. and G. K. BM. Methods for Identifying Differentially Expressed Genes : An Empirical Comparison. *Journal of Biometrics & Biostatistics*, 06(05), dec 2015. [27](#)
- [54] A. A. Hagberg, D. A. Schult, and P. J. Swart. Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference (SciPy2008)*, pages 11–15, Pasadena, CA USA, Aug. 2008. [51](#), [69](#)
- [55] M. Hallek, C. Neumann, M. Schäffer, S. Danhauser-Riedl, N. von Bubnoff, G. de Vos, B. J. Druker, K. Yasukawa, J. D. Griffin, and B. Emmerich. Signal transduction of interleukin-6 involves tyrosine phosphorylation of multiple cytosolic proteins and activation of Src-family kinases Fyn, Hck, and Lyn in multiple myeloma cell lines. *Experimental hematology*, 25(13) :1367–77, dec 1997. [47](#)
- [56] H. Han, H. Shim, D. Shin, J. E. Shim, Y. Ko, J. Shin, H. Kim, A. Cho, E. Kim, T. Lee, et al. Trustrust : a reference database of human transcriptional regulatory interactions. *Scientific reports*, 5 :11432, 2015. [13](#)
- [57] J. Herbrand. Recherches sur la théorie de la démonstration. [23](#)
- [58] I. Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011) :931–945, oct 2004. [10](#)
- [59] T. Ideker, O. Ozier, B. Schwikowski, and A. F. Siegel. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics (Oxford, England)*, 18 Suppl 1 :S233–40, Jan. 2002. [36](#)
- [60] H. Ishikawa. Requirements of src family kinase activity associated with CD45 for myeloma cell proliferation by interleukin-6. *Blood*, 99(6) :2172–2178, mar 2002. [47](#)
- [61] N. Jamshidi and B. O. Palsson. Using in silico models to simulate dual perturbation experiments : procedure development and interpretation of outcomes. *BMC Systems Biology*, 3(1) :44, apr 2009. [87](#)
- [62] J. B. Johnston, S. Navaratnam, M. W. Pitz, J. M. Maniate, E. Wiechec, H. Baust, J. Gingerich, G. P. Skliris, L. C. Murphy, and M. Los. Targeting the EGFR pathway for cancer therapy. *Current medicinal chemistry*, 13(29) :3483–3492, jan 2006. [47](#)
- [63] A. Kalff and A. Spencer. The t(4;14) translocation and FGFR3 overexpression in multiple myeloma : prognostic implications and current clinical strategies. *Blood cancer journal*, 2(9) :e89, sep 2012. [50](#)

- [64] M. Kanehisa and S. Goto. KEGG : kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1) :27–30, jan 2000. [13](#)
- [65] K. Kapur, Y. Xing, Z. Ouyang, and W. H. Wong. Exon arrays provide accurate assessments of gene expression. *Genome biology*, 8(5) :R82, 2007. [11](#)
- [66] J. J. Keats, R. Fonseca, M. Chesi, R. Schop, A. Baker, W.-J. Chng, S. Van Wier, R. Tiedemann, C.-X. Shi, M. Sebag, et al. Promiscuous mutations activate the noncanonical nf-kb pathway in multiple myeloma. *Cancer cell*, 12(2) :131–144, 2007. [32](#)
- [67] P. Khatri, M. Sirota, and A. J. Butte. Ten years of pathway analysis : current approaches and outstanding challenges. *PLoS computational biology*, 8(2) :e1002375, jan 2012. [33](#)
- [68] A. Kittas, A. Delobelle, S. Schmitt, K. Breuhahn, C. Guziolowski, and N. Grabe. Directed random walks and constraint programming reveal active pathways in hepatocyte growth factor signaling. *FEBS Journal*, 283(2) :350–360, jan 2016. [37](#), [41](#), [43](#)
- [69] B. Klein. Positioning NK-kappaB in multiple myeloma. *Blood*, 115(17) :3422–4, Apr. 2010. [44](#), [73](#)
- [70] K. Komurov, S. Dursun, S. Erdin, and P. T. Ram. NetWalker : a contextual network analysis tool for functional genomics. *BMC genomics*, 13(1) :282, Jan. 2012. [36](#), [37](#)
- [71] S. W. Kong, W. T. Pu, and P. J. Park. A multivariate approach for integrating genome-wide expression data and biological knowledge. *Bioinformatics (Oxford, England)*, 22(19) :2373–80, oct 2006. [36](#)
- [72] M. Kutmon, A. Riutta, N. Nunes, K. Hanspers, E. Willighagen, A. Bohler, J. Mélius, A. Waagmeester, S. Sinha, R. Miller, S. L. Coort, E. Cirillo, B. Smeets, C. Evelo, and A. R. Pico. WikiPathways : capturing the full diversity of pathway knowledge. *Nucleic Acids Research*, 44(D1) :D488–D494, jan 2016. [15](#)
- [73] R. A. Kyle. Multiple myeloma : how did it begin ? *Mayo Clinic proceedings*, 69(7) :680–3, jul 1994. [30](#)
- [74] Lederberg Joshua. 'Ome Sweet 'Omics– A Genealogical Treasury of Words | The Scientist Magazine®, apr 2001. [12](#)
- [75] C. Lefebvre, P. Rajbhandari, M. J. Alvarez, P. Bandaru, W. K. Lim, M. Sato, K. Wang, P. Sumazin, M. Kustagi, B. C. Bisikirska, K. Basso, P. Beltrao, N. Krogan, J. Gautier, R. Dalla-Favera, and A. Califano. A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers. *Molecular systems biology*, 6 :377, jun 2010. [33](#), [35](#), [36](#)
- [76] T. Lenoir and E. Giannella. The emergence and diffusion of DNA microarray technology. *Journal of biomedical discovery and collaboration*, 1(1) :11, jan 2006. [11](#)

- [77] Y. Li, L. Liu, X. Bai, H. Cai, W. Ji, D. Guo, and Y. Zhu. Comparative study of discretization methods of microarray data for inferring transcriptional regulatory networks. *BMC bioinformatics*, 11(1) :520, jan 2010. [29](#)
- [78] V. Lifschitz. What Is Answer Set Programming ? [24](#)
- [79] W. Liu, C. Li, Y. Xu, H. Yang, Q. Yao, J. Han, D. Shang, C. Zhang, F. Su, X. Li, Y. Xiao, F. Zhang, M. Dai, and X. Li. Topologically inferring risk-active pathways toward precise cancer classification by directed random walk. *Bioinformatics (Oxford, England)*, 29(17) :2169–77, Sept. 2013. [36](#), [37](#)
- [80] S. Lloyd. Least square quantization in pcm. bell telephone laboratories paper. published in journal much later : Lloyd, sp : Least squares quantization in pcm. *IEEE Trans. Inform. Theor.(1957/1982)*. [18](#)
- [81] S. Lub, K. Maes, E. Menu, E. De Bruyne, K. Vanderkerken, and E. Van Valckenborgh. Novel strategies to target the ubiquitin proteasome system in multiple myeloma. *Oncotarget*, 7(6) :6521–37, feb 2016. [30](#)
- [82] S. Madeira and A. Oliveira. An evaluation of discretization methods for non-supervised analysis of time-series gene expression data. instituto de engenharia de sistemas e computadores investigacao e desenvolvimento. Technical report, Technical Report 42, 2005. [29](#)
- [83] F. Magrangeas, H. Avet-Loiseau, W. Gouraud, L. Lodé, O. Decaux, P. Godmer, L. Garderet, L. Voillat, T. Facon, a. M. Stoppa, G. Marit, C. Hulin, P. Casassus, M. Tiab, E. Voog, E. Randriamalala, K. C. Anderson, P. Moreau, N. C. Munshi, and S. Minvielle. Minor clone provides a reservoir for relapse in multiple myeloma. *Leukemia*, 27(2) :473–81, feb 2013. [31](#)
- [84] K. Mahtouk, D. Hose, T. Rème, J. De Vos, M. Jourdan, J. Moreaux, G. Fiol, M. Raab, E. Jourdan, V. Grau, M. Moos, H. Goldschmidt, M. Baudard, J. F. Rossi, F. W. Cremer, and B. Klein. Expression of EGF-family receptors and amphiregulin in multiple myeloma. Amphiregulin is a growth factor for myeloma cells. *Oncogene*, 24(21) :3512–3524, may 2005. [47](#)
- [85] K. Mahtouk, M. Jourdan, J. De Vos, C. Hertogh, G. Fiol, E. Jourdan, J.-F. Rossi, and B. Klein. An inhibitor of the EGF receptor family blocks myeloma cell growth factor activity of HB-EGF and potentiates dexamethasone or anti-IL-6 antibody-induced apoptosis. *Blood*, 103(5) :1829–37, mar 2004. [47](#)
- [86] F. Martin, A. Sewer, M. Talikka, Y. Xiang, J. Hoeng, and M. C. Peitsch. Quantification of biological network perturbations for mechanistic insight and diagnostics using two-layer causal models. *BMC bioinformatics*, 15(1) :238, jan 2014. [34](#), [36](#), [37](#)
- [87] A. Marusyk, V. Almendro, and K. Polyak. Intra-tumour heterogeneity : a looking glass for cancer ? *Nature reviews. Cancer*, 12(5) :323–34, apr 2012. [31](#), [90](#)

- [88] V. Marx. Biology : The big challenges of big data. *Nature*, 498(7453) :255–260, jun 2013. [13](#), [26](#)
- [89] M. N. McCall, H. a. Jaffee, and R. a. Irizarry. fRMA ST : frozen robust multiarray analysis for Affymetrix Exon and Gene ST arrays. *Bioinformatics (Oxford, England)*, 28(23) :3153–4, dec 2012. [44](#)
- [90] M. N. McCall, H. A. Jaffee, S. J. Zelisko, N. Sinha, G. Hooiveld, R. A. Irizarry, and M. J. Zilliox. The Gene Expression Barcode 3.0 : improved data processing and mining tools. *Nucleic acids research*, 42(Database issue) :D938–43, jan 2014. [44](#)
- [91] D. J. McCarthy and G. K. Smyth. Testing significance relative to a fold-change threshold is a TREAT. *Bioinformatics (Oxford, England)*, 25(6) :765–71, mar 2009. [27](#)
- [92] H. Mi, A. Muruganujan, and P. D. Thomas. PANTHER in 2013 : modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic acids research*, 41(Database issue) :D377–86, jan 2013. [33](#), [34](#), [76](#)
- [93] B. Miannay, S. Minvielle, O. Roux, P. Drouin, H. Avet-Loiseau, C. Guérin-Charbonnel, W. Gouraud, M. Attal, T. Facon, N. C. Munshi, P. Moreau, L. Campion, F. Magrangeas, and C. Guziolowski. Logic programming reveals alteration of key transcription factors in multiple myeloma. *Scientific Reports*, 7(1) :9257, 2017. [15](#), [86](#), [105](#)
- [94] B. Miannay, S. Minvielle, O. Roux, F. Magrangeas, and C. Guziolowski. Constraints On Signaling Networks Logic Reveal Functional Subgraphs On Multiple Myeloma OMIC Data. In *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics - ACM-BCB '17*, pages 768–769, New York, New York, USA, 2017. ACM Press. [15](#), [86](#), [105](#)
- [95] G. J. Morgan, B. A. Walker, and F. E. Davies. The genetic architecture of multiple myeloma. *Nature reviews. Cancer*, 12(5) :335–48, may 2012. [30](#), [31](#), [32](#), [50](#)
- [96] J. N. Morgan and J. A. Sonquist. Problems in the analysis of survey data, and a proposal. *Journal of the American statistical association*, 58(302) :415–434, 1963. [19](#)
- [97] J. H. Morris, L. Apeltsin, A. M. Newman, J. Baumbach, T. Wittkop, G. Su, G. D. Bader, and T. E. Ferrin. clusterMaker : a multi-algorithm clustering plugin for Cytoscape. *BMC bioinformatics*, 12 :436, nov 2011. [78](#)
- [98] F. Muhlenbach and R. Rakotomalala. Discretization for continuous attributes. In *Encyclopedia of Data Warehousing and Mining*, pages 397–402. IGI Global, 2005. [28](#)
- [99] U. Nagalakshmi, Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein, and M. Snyder. The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. *Science*, 320(5881) :1344–1349, jun 2008. [11](#)
- [100] T. Nepusz, H. Yu, and A. Paccanaro. Detecting overlapping protein complexes in protein-protein interaction networks. *Nature methods*, 9(5) :471–472, 2012. [78](#)

- [101] J. R. Nevins. The Rb/E2F pathway and cancer. *Human molecular genetics*, 10(7) :699–703, apr 2001. [49](#)
- [102] J. K. Nicholson. Reviewers peering from under a pile of ‘omics’ data. *Nature*, 440(7087) :992–992, apr 2006. [12](#)
- [103] R. Niesvizky, A. Z. Badros, L. J. Costa, S. A. Ely, S. B. Singhal, E. A. Stadtmauer, N. A. Haideri, A. Yacoub, G. Hess, S. Lentzsch, I. Spicka, A. A. Chanan-Khan, M. S. Raab, S. Tarantolo, R. Vij, J. A. Zonder, X. Huang, D. Jayabalan, M. Di Liberto, X. Huang, Y. Jiang, S. T. Kim, S. Randolph, and S. Chen-Kiang. Phase 1/2 study of cyclin-dependent kinase (CDK)4/6 inhibitor palbociclib (PD-0332991) with bortezomib and dexamethasone in relapsed/refractory multiple myeloma. *Leukemia & Lymphoma*, 56(12) :3320–3328, dec 2015. [32](#), [49](#)
- [104] H. Osaki, A. Sasaki, E. Nakazono-Nagaoka, N. Ota, and R. Nakaune. Genome segments encoding capsid protein-like variants of *Pyrus pyrifolia* cryptic virus. *Virus Research*, jul 2017. [10](#)
- [105] G. S. Pall and A. J. Hamilton. Improved northern blot method for enhanced detection of small RNA. *Nature Protocols*, 3(6) :1077–1084, jun 2008. [10](#)
- [106] C. Pan, X. Yan, H. Li, L. Huang, M. Yin, Y. Yang, R. Gao, L. Hong, Y. Ma, C. Shi, H. Qin, P. Zhang, C. Pan, X. Yan, H. Li, L. Huang, M. Yin, Y. Yang, R. Gao, L. Hong, Y. Ma, C. Shi, H. Qin, and P. Zhang. Systematic literature review and clinical validation of circulating microRNAs as diagnostic biomarkers for colorectal cancer. *Oncotarget*, 5(0), jul 2017. [10](#)
- [107] C. M. Perou, S. S. Jeffrey, M. van de Rijn, C. A. Rees, M. B. Eisen, D. T. Ross, A. Pergamenschikov, C. F. Williams, S. X. Zhu, J. C. Lee, D. Lashkari, D. Shalon, P. O. Brown, and D. Botstein. Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proceedings of the National Academy of Sciences of the United States of America*, 96(16) :9212–7, aug 1999. [11](#)
- [108] R. T. Pillich, J. Chen, V. Rynkov, D. Welker, and D. Pratt. NDEx : A Community Resource for Sharing and Publishing of Biological Networks. In *Methods in molecular biology (Clifton, N.J.)*, volume 1558, pages 271–301. 2017. [14](#)
- [109] K. Podar, M. S. Raab, G. Tonon, M. Sattler, D. Barilà, J. Zhang, Y.-T. Tai, H. Yasui, N. Raje, R. A. DePinho, T. Hideshima, D. Chauhan, and K. C. Anderson. Up-regulation of c-Jun inhibits proliferation and induces apoptosis via caspase-triggered c-Abl cleavage in human multiple myeloma. *Cancer research*, 67(4) :1680–8, feb 2007. [47](#), [52](#)
- [110] J. Quackenbush. Microarray data normalization and transformation. *Nature Genetics*, 32(Supp) :496–501, dec 2002. [17](#)
- [111] J. Quinlan. Simplifying decision trees. *International Journal of Man-Machine Studies*, 27(3) :221–234, sep 1987. [46](#)

- [112] R Core Team. *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. [51](#), [69](#)
- [113] S. V. Rajkumar. Multiple myeloma : 2016 update on diagnosis, risk-stratification, and management. *American Journal of Hematology*, 91(7) :719–734, jul 2016. [31](#), [50](#)
- [114] N. U. Rashid, A. S. Sperling, N. Bolli, D. C. Wedge, P. Van Loo, Y.-T. Tai, M. A. Shamas, M. Fulciniti, M. K. Samur, P. G. Richardson, F. Magrangeas, S. Minvielle, P. A. Futreal, K. C. Anderson, H. Avet-Loiseau, P. J. Campbell, G. Parmigiani, and N. C. Munshi. Differential and limited expression of mutant alleles in multiple myeloma. *Blood*, 124(20), 2014. [32](#), [89](#)
- [115] A. Razi, F. Afghah, S. Singh, and V. Varadan. Network-Based Enriched Gene Subnetwork Identification : A Game-Theoretic Approach. *Biomed Eng Comput Biol*, 7(Suppl 2) :1–14, 2016. [36](#), [37](#), [63](#)
- [116] J. Saez-Rodriguez, L. G. Alexopoulos, J. Epperlein, R. Samaga, D. a. Lauffenburger, S. Klamt, and P. K. Sorger. Discrete logic modelling as a means to link protein signalling networks with functional analysis of mammalian signal transduction. *Molecular systems biology*, 5(331) :331, Jan. 2009. [44](#)
- [117] M. N. Saha, H. Jiang, Y. Yang, X. Zhu, X. Wang, A. D. Schimmer, L. Qiu, and H. Chang. Targeting p53 via JNK pathway : a novel role of RITA for apoptotic signaling in multiple myeloma. *PLoS one*, 7(1) :e30215, jan 2012. [47](#), [52](#)
- [118] F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12) :5463–7, dec 1977. [11](#)
- [119] M. P. Sawicki, G. Samara, M. Hurwitz, and E. Passaro. Human Genome Project. *The American Journal of Surgery*, 165(2) :258–264, feb 1993. [11](#)
- [120] C. F. Schaefer, K. Anthony, S. Krupa, J. Buchoff, M. Day, T. Hannay, and K. H. Buetow. PID : the Pathway Interaction Database. *Nucleic acids research*, 37(Database issue) :D674–9, Jan. 2009. [13](#), [43](#), [66](#), [73](#)
- [121] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown. Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. *Science*, 270(5235) :467–470, oct 1995. [10](#)
- [122] E. M. Sewify, O. A. Afifi, E. Mosad, A. H. Zaki, and S. A. El Gammal. Cyclin D1 Amplification in Multiple Myeloma Is Associated With Multidrug Resistance Expression. *Clinical Lymphoma Myeloma and Leukemia*, 14(3) :215–222, jun 2014. [30](#)
- [123] K. Siklenka, S. Erkek, M. Godmann, R. Lambrot, S. McGraw, C. Laffleur, T. Cohen, J. Xia, M. Suderman, M. Hallett, J. Trasler, A. H. F. M. Peters, and S. Kimmins. Disruption of histone methylation in developing sperm impairs offspring health transgenerationally. *Science*, 2015. [10](#)

- [124] G. K. Smyth. Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1) :1–25, jan 2004. [27](#)
- [125] H. Steinhaus. Sur la division des corps materiels en parties. *bull. acad. polon. sci.*, c1. iii vol iv : 801-804. 1956. [18](#)
- [126] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis : a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43) :15545–50, Oct. 2005. [33](#), [35](#), [36](#)
- [127] The Gene Ontology Consortium. Gene Ontology : tool for the unification of biology. *Nature Genetics*, 25(1) :25–29, may 2000. [13](#), [33](#), [76](#)
- [128] The UniProt Consortium. Activities at the Universal Protein Resource (UniProt). *Nucleic acids research*, 42(Database issue) :D191–8, jan 2014. [13](#)
- [129] S. Thiele, L. Cerone, J. Saez-Rodriguez, A. Siegel, C. Guziolowski, and S. Klamt. Extended notions of sign consistency to relate experimental data to signaling and regulatory network topologies. *BMC bioinformatics*, 16(1) :345, jan 2015. [29](#), [36](#), [37](#), [51](#), [52](#), [64](#)
- [130] C. Touriol, S. Bornes, S. Bonnal, S. Audigier, H. Prats, A.-C. Prats, and S. Vagner. Generation of protein isoform diversity by alternative initiation of translation at non-AUG codons. *Biology of the Cell*, 95(3-4) :169–178, may 2003. [9](#)
- [131] D. Turei, T. Korcsmaros, and J. Saez-Rodriguez. OmniPath : guidelines and gateway for literature-curated signaling pathway resources. *Nat Meth*, 13(12) :966–967, dec 2016. [15](#)
- [132] V. Tusher, R. Tibshirani, and C. Chu. Significance analysis of microarrays applied to ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98 :5116–5121, 2001. [27](#)
- [133] S. Uddin, A. R. Hussain, M. Ahmed, K. Siddiqui, F. Al-Dayel, P. Bavi, and K. S. Al-Kuraya. Overexpression of FoxM1 offers a promising therapeutic target in diffuse large B-cell lymphoma. *Haematologica*, 97(7) :1092–100, jul 2012. [32](#), [47](#), [48](#)
- [134] B. A. Walker, E. M. Boyle, C. P. Wardell, A. Murison, D. B. Begum, N. M. Dahir, P. Z. Proszek, D. C. Johnson, M. F. Kaiser, L. Melchor, L. I. Aronson, M. Scales, C. Pawlyn, F. Mirabella, J. R. Jones, A. Brioli, A. Mikulasova, D. A. Cairns, W. M. Gregory, A. Quartilho, M. T. Drayson, N. Russell, G. Cook, G. H. Jackson, X. Leleu, F. E. Davies, and G. J. Morgan. Mutational Spectrum, Copy Number Changes, and Outcome : Results of a Sequencing Study of Patients With Newly Diagnosed Myeloma. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 33(33) :3911–20, nov 2015. [32](#)

- [135] A. J. Waxman, P. J. Mink, S. S. Devesa, W. F. Anderson, B. M. Weiss, S. Y. Kristinsson, K. A. McGlynn, and O. Landgren. Racial disparities in incidence and outcome in multiple myeloma : a population-based study. *Blood*, 116(25), 2010. 31
- [136] J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. M. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, J. M. Stuart, C. G. A. R. Network, et al. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10) :1113–1120, 2013. 13
- [137] E. W. Weisstein. Bonferroni correction. 2004. Disponible sur le site : <http://math-world.wolfram.com/BonferroniCorrection.html>. 27
- [138] Wikipédia. Acide désoxyribonucléique — wikipédia, l’encyclopédie libre, 2017. [En ligne ; Page disponible le 3-août-2017]. 8
- [139] Wikipédia. K-moyennes — wikipédia, l’encyclopédie libre, 2017. [En ligne ; Page disponible le 1-août-2017]. 19
- [140] Wikipédia. Puce à adn — wikipédia, l’encyclopédie libre, 2017. [En ligne ; Page disponible le 2-août-2017]. 11
- [141] E. Wingender. The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. *Briefings in Bioinformatics*, 9(4) :326–332, mar 2008. 10
- [142] F. H. Xu, S. Sharma, a. Gardner, Y. Tu, a. Raitano, C. Sawyers, and a. Lichtenstein. Interleukin-6-induced inhibition of multiple myeloma cell apoptosis : support for the hypothesis that protection is mediated via inhibition of the JNK/SAPK pathway. *Blood*, 92(1) :241–251, jul 1998. 47
- [143] Ö. Yaveroğlu, T. Milenković, and N. Pržulj. Proper evaluation of alignment-free network comparison methods. *Bioinformatics*, 2015. 36
- [144] F. Zhan, Y. Huang, S. Colla, J. P. Stewart, I. Hanamura, S. Gupta, J. Epstein, S. Yaccoby, J. Sawyer, B. Burington, E. Anaissie, K. Hollmig, M. Pineda-Roman, G. Tricot, F. Van Rhee, R. Walker, M. Zangari, J. Crowley, B. Barlogie, and J. D. Shaughnessy. The molecular classification of multiple myeloma. *Blood*, 108(6) :2020–2028, sep 2006. 50
- [145] Y. Zhang, D. Liu, X. Chen, J. Li, L. Li, Z. Bian, F. Sun, J. Lu, Y. Yin, X. Cai, Q. Sun, K. Wang, Y. Ba, Q. Wang, D. Wang, J. Yang, P. Liu, T. Xu, Q. Yan, J. Zhang, K. Zen, and C.-Y. Zhang. Secreted Monocytic miR-150 Enhances Targeted Endothelial Cell Migration. *Molecular Cell*, 39(1) :133–144, jul 2010. 10

Annexe

.1 Production et communication scientifique

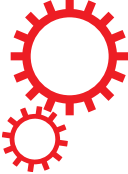
Ce paragraphe vise à présenter sous forme de liste les publications, posters, présentations et les autres activités de diffusions scientifique qui ont pu être faites durant cette thèse

- Journaux avec comité de relecture :
 - Article dans Scientific Reports sur la modélisation par coloration de graphes pour le MM (Chapitre 3) [93] (Voir annexe .2)
 - Article en cours de soumission sur l’exploration des colorations parfaites et son application sur le myélome multiple (Chapitre 4).
- Actes publiés de conférence
 - 2017 : *Workshop* CNB-MAC, Boston, USA [94].
- Présentations orales
 - 14 décembre 2015 : Journées scientifiques de la bioinformatique de Nantes, Nantes, France (goo.gl/vY1qNF).
 - 30 juin 2016 : Séminaire Bioss, Lyon, France (goo.gl/SVgd4w).
 - 8 juin 2017 : Séminaire Nohmad, Nantes, France (goo.gl/h4wgHn).
 - 20 août 2017 : *Workshop* CNB-MAC, Boston, USA (goo.gl/uPTc2d).
- Posters
 - 9-14 novembre 2014 : Conférence RECOMB/ICSB, San Diego, USA (goo.gl/ZJfowG) [14].

- 16-18 septembre 2015 : Conférence CMSB, Nantes, France (goo.gl/Fqs8Yi) [13].
- 12-15 novembre 2016 : Conférence EMBO conference : From Functional Genomics to Systems Biology, Heidelberg, Allemagne (goo.gl/TqUqdg) [11].
- Enseignement et vulgarisation scientifique
 - Enseignement “Algorithmique et programmation” (TD + TP)
 - Enseignement “Base de données” (TP).
 - 2016 : Encadrement de stage sur la discrétisation des données.
 - 2015 : Entretien en journal grand public (ouest France) (goo.gl/mUqMWC).
 - 2015 : Participation à la fête de la science : "Tête à tête, jeunes et chercheurs".
 - 2016 : Présentation de la bio-informatique au lycée Savary de Mauléon (goo.gl/FvVWbA).

.2 Article publié dans Scientific Reports

SCIENTIFIC REPORTS



OPEN

Logic programming reveals alteration of key transcription factors in multiple myeloma

Bertrand Miannay^{1,2}, Stéphane Minvielle^{2,3}, Olivier Roux¹, Pierre Drouin¹, Hervé Avet-Loiseau⁴, Catherine Guérin-Charbonnel^{2,5}, Wilfried Gouraud^{2,5}, Michel Attal⁶, Thierry Facon⁷, Nikhil C Munshi^{8,9}, Philippe Moreau^{2,3}, Loïc Campion^{2,5}, Florence Magrangeas^{2,3} & Carito Guziolowski¹

Innovative approaches combining regulatory networks (RN) and genomic data are needed to extract biological information for a better understanding of diseases, such as cancer, by improving the identification of entities and thereby leading to potential new therapeutic avenues. In this study, we confronted an automatically generated RN with gene expression profiles (GEP) from a cohort of multiple myeloma (MM) patients and normal individuals using global reasoning on the RN causality to identify key-nodes. We modeled each patient by his or her GEP, the RN and the possible automatically detected repairs needed to establish a coherent flow of the information that explains the logic of the GEP. These repairs could represent cancer mutations leading to GEP variability. With this reasoning, unmeasured protein states can be inferred, and we can simulate the impact of a protein perturbation on the RN behavior to identify therapeutic targets. We showed that JUN/FOS and FOXM1 activities are altered in almost all MM patients and identified two survival markers for MM patients. Our results suggest that JUN/FOS-activation has a strong impact on the RN in view of the whole GEP, whereas FOXM1-activation could be an interesting way to perturb an MM subgroup identified by our method.

Multiple myeloma (MM) is a neoplasm of plasma cells with an incidence rate of approximately 5/100,000 in Europe. The median survival of MM patients has improved substantially over the past decade. Owing to the establishment of high-dose therapy followed by autologous stem cell transplantation as a routine procedure, significant improvements in supportive care strategies, and the introduction and widespread use of the immunomodulatory drugs thalidomide and lenalidomide, and the proteasome inhibitor bortezomib. Nevertheless, almost all MM patients ultimately relapse, and new drugs and new combinations for the treatment of MM are warranted. MM is a heterogeneous disease at both the clinical and molecular levels. Recent large scale genomics analysis based on the landscape of copy-number alterations and on whole exome sequencing have revealed the hallmarks of genetic changes in MM such as hyperdiploidy, translocations involving the IgH locus, and mutations in the RAS/MAP and NF- κ B pathways and in TP53¹. These genetic changes as well as gene-expression profiling (GEP) have been widely used in the molecular classification of newly diagnosed patients to define diagnostic entities and identify promising new therapeutic targets²⁻⁷. However, at present a standard of classification based on subgroups that could be targeted therapeutically is still being debated. Clearly, there is a need for innovative tools to improve the identification of the prognostically relevant entities, clinically and biologically, in newly diagnosed MM patients. It is tempting to use the mutational spectrum based on whole-exome sequencing as a gold standard; however we have previously shown that a large number of exome mutant alleles are not expressed clinically or biologically⁸. In addition, exome sequencing may miss potential driver mutations in the non coding regulatory elements known to affect enhancer activity, which thereby affect the transcriptional program⁹; therefore GEP remains a tool of choice.

¹LS2N, UMR 6004, École Centrale de Nantes, Nantes, France. ²CRCINA, INSERM, CNRS, Université d'Angers, Université de Nantes, Nantes, France. ³CHU de Nantes, Nantes, France. ⁴Unit for Genomics in Myeloma, IUC-Oncopole; and, CRCT INSERM 1037, Toulouse, France. ⁵Institut de Cancérologie de l'Ouest, Nantes, France. ⁶Department of Hematology, IUC, Toulouse, France. ⁷Department of Hematology, CHU, Lille, France. ⁸Lebown Institute of Myeloma Therapeutics and Jerome Lipper Multiple Myeloma Center, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA, 02115, USA. ⁹Boston Veterans Administration Healthcare System, West Roxbury, MA, 02132, USA. Correspondence and requests for materials should be addressed to F.M. (email: Florence.Magrangeas@univ-nantes.fr) or C.G. (email: Carito.Guziolowski@ls2n.fr)

However, GEP alone is limited and must be integrated with innovative approaches that use biological regulatory networks to extract biological information relative to gene expression datasets to provide significant clues about the etiology of myeloma.

During the past decade, many methods of so-called *pathway analysis* or *active pathways detection* have been developed. These methods use as a knowledge base a biological pathway or regulatory network, that compiles a series of molecular phenomena that lead to activation (or inhibition) of gene expression, a cell product such as a hormone, or a physical modification of the cell. Regulatory network information is currently available through databases such as Gene Ontology (GO)¹⁰, the Kyoto Encyclopedia of Genes and Genomes (KEGG)¹¹, the Pathway Interaction Database (PID)¹², Wikipathway¹³, Transfac¹⁴, and Causal Biological Networks (CBN)¹⁵. The main objective of pathway analysis methods is to confront or integrate GEP data with regulatory networks or pathways to distinguish two or more classes of cells (e.g. healthy vs ill) from GEP data by inferring a specific signature for each class. We can identify three principal categories of approaches that have been used to associate GEP with specific pathways¹⁶.

The Over-Representation Analysis (ORA) group of approaches^{17, 18} includes approaches that are based on differentially expressed (DE) genes. These approaches score single pathways based on the proportion of DE genes (identified with statistical tests or with a threshold) contained in each pathway. In most cases, these methods use a hyper-geometric test¹⁷ to score each pathway. Moreover, the majority of ORA approaches that use functional annotation (GO) or pathway maps (KEGG) consider the consequences of the DE genes (leading to the differential expression of proteins) in the associations between gene and pathway. Martin *et al.*¹⁹ called this type of reasoning *forward assumption* compared to the *backward assumption*¹⁸, which considers the causes of those DE genes in the gene-pathway association.

The Functional Class Scoring (FCS) group of approaches uses the full datasets without any pre-selection, allowing integration of the effects of low gene expression variations in the identification of the pathways involved. FCS approaches can use forward^{20, 21} or backward^{22, 23} reasoning. Although these methods improve the problem of genes selection, the pathways in which individual genes are involved are still studied independently. Moreover, the position of the genes in the topology is not used in the analysis.

The Pathway Topology (PT) approaches are very similar to the FCS approaches, but in addition, they score genes according to the pathways to which they belong. Whereas some of these approaches only include interactions between genes^{24–27}, others consider different types of relationships between genes^{19, 28}, generally activation and inhibition. The majority of methods study each pathway independently. Within this group, we can also identify methods that use both forward^{24–27} and backward^{19, 28} reasoning.

In this work, we propose to integrate the GEPs obtained from myeloma cells (MC) of 602 MM patients and from normal plasma cells (NPC) of 9 healthy donors with the whole compendium of the PID-NCI public pathway repository so as to better understand the mechanisms of plasma cell carcinogenesis. To integrate this data, we first automatically build a directed (and labeled) graph using the whole compendium of the PID-NCI public pathway repository. This graph connects signaling pathways to the transcription of the genes in the GEP dataset. We then integrate the graph with the expression data by reasoning on its logic using IGGY²⁹, a tool based on logic programming (Answer Set Programming) that confronts a node coloring (GEP) with labeled and directed graphs. Our combined approach could be considered to fall within the PT category since it takes into account the causality and activation/inhibition logic of graph edges. However, unlike previous methods, it uses a global logic to analyze experimental and pathway data. In this formalism, both forward and backward modes are included as reasoning modes (causes-consequences). IGGY allows us to check the consistency of the information and to generate predictions based upon automatic repairs for upstream non-measured species. It uses DE data as well as the identically expressed genes across classes (invariant genes) in its analysis. The proposed method does not correlate protein activation with gene expression; the two entities are identified separately in the graph. The non-measured protein activations necessary to satisfy the GEP according to the entire pathway database topology are used later to propose a signature for each dataset profile. This global signature can be used to characterize the dataset classes. Moreover, our model also allows us to *in silico* quantify the effect of perturbations on this global pathway for each single patient. We show how this type of method, which combines large-scale information in terms of number of patients, the complete GEP, and the entire compendium database, can be applied to identify new specificities of MM disease compared to normal cells. As a result, we inferred information on the states of specific proteins in the cell that may cause these disorders, and we identified specific markers of MC compared to NPC that can be used to identify survival markers. Furthermore, these markers can be studied as therapeutic targets because of their over-representation and their impact on the involved pathways.

Materials and Methods

Data. *Experimental Procedures.* Plasma cells were isolated from the bone marrow of 602 newly diagnosed cases of MM. The samples were obtained during standard diagnostic procedures conducted at the Intergroupe Francophone du Myélome (IFM) centers. The subjects included patients younger than 65 years of age who were enrolled in either the IFM 2005–01 trial (n = 311) or the IFM-2007-02 (n = 128) trial, older patients enrolled in the IFM-2007-01/Multiple Myeloma 020 trial (n = 76) and 9 normal donors. The experiments were undertaken with the understanding and written informed consent of each subject. Plasma cell purification was performed as previously described³⁰. Purified plasma cells were frozen at –80 °C in lysis buffer. Approval for this study was obtained from the University Hospital of Nantes. The study fulfilled the requirements of the Declaration of Helsinki.

Gene expression profiling. RNA was extracted using the AllPrep DNA/RNA MiniKit or the RNeasy Micro kit (QIAGEN, Valencia, CA, USA) in accordance with the manufacturer's instructions. RNA quality and quantity were assessed using Agilent 2100 Bioanalyzer (Agilent, Palo Alto, CA, USA) and a Nanodrop Spectrophotometer

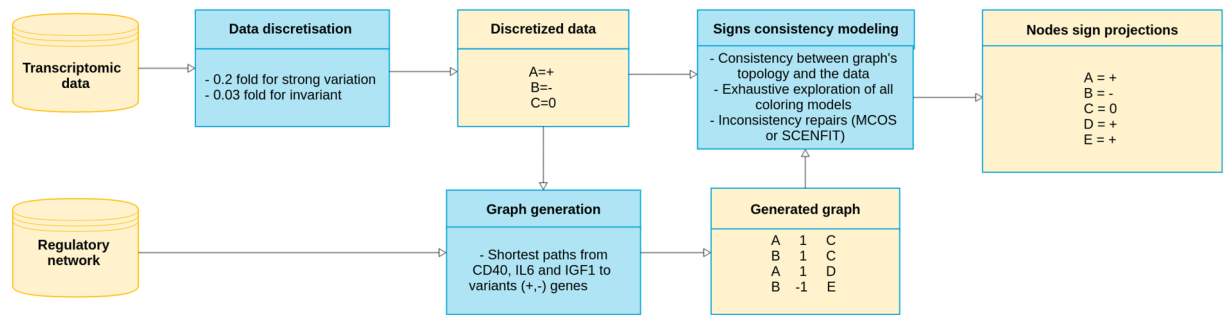


Figure 1. Overview of the sign consistency modeling framework.

(NanoDrop Technologies, DE, USA), respectively. MM samples from which 50 ng of total RNA was available were processed according to manufacturer's instructions (NuGEN, San Carlos, CA, USA) before labeling and hybridization onto an Affymetrix Human Exon1.0 chip according to the manufacturer's instructions (Affymetrix, Santa Clara, CA). Data were analyzed and \log_2 normalized with Expression Console Affymetrix software v1.1 using an RMA algorithm.

Data discretization. Since our *graph coloring model* requires invariant genes, classical discretization methods cannot be used in our study. To identify over-/underexpressed and invariant gene expression for each profile, we used two thresholds: k_1 for invariant genes and k_2 for variant genes. For each gene g , we computed a vector, p_i^g , composed of its differential expression, p_i^g , in each dataset expression profile i . A profile i refers to each of the 611 cells of type MC or NPC considered in this study. p_i^g values were computed by subtracting the mean expression of g in the NPC set from its gene expression level in MC (Supplementary Material, Figure S1). We then discretized the values of p_i^g using two thresholds k_1 and k_2 .

If $p_i^g > k_2$, g was considered over-expressed for i ;
 if $p_i^g < -k_2$, g was considered under-expressed for i ;
 and if $-k_1 < p_i^g < k_1$, g was considered invariant for i .

By choosing different combinations of values for k_1 and k_2 (see Supplementary Material), we obtained 150 sets of vectors that contain the discrete overexpressed (+), underexpressed (-), or invariant (0) values for all the genes expressed in each dataset. We discarded combinations of values leading to p^g vectors with a sign proportion greater than 50%. Each $k_1 - k_2$ combination was used to test the precision of our approach. We did this 100 times by using 50% of the discretized $\{+, -, 0\}$ genes' expression to predict the other 50% for each MC dataset, after which we comparing the measured and predicted data using a precision matrix (Supplementary Material, Table S1). The thresholds leading to the best precision of 43% (IC 95%: $\pm 3\%$) were $k_1 = 0.03$ and $k_2 = 0.2$; these thresholds were used in the remainder of the study to select the variant and invariant genes. In the Supplementary Material, Figure S2, we show the precision obtained for all of the threshold combinations that were selected. Since our discretization method fixes the same thresholds for all genes across all profiles, we also used K-means to discover gene-specific thresholds. However, the precision of K-means methods (for $k = 3$) was lower than that obtained using the selected thresholds (see Supplementary Material, Figure S3). In the same way, to demonstrate the interest of using invariant genes, we computed the precision of recovering 50% of the data using a two-signs model that receives input data and predicts only over- and underexpressed values. The computed precision was 48%. Note that the two-signs model has a precision closer to a random precision distribution (50%), whereas the precision obtained using a three-signs model is farther from the random precision (33%).

Graph generation. We used the 2012 version of the complete pathways database PID-NCI (Pathway Interaction Database)¹² and downloaded it in PID-XML format. This database is specialized to include regulatory pathways involved in cancer. The complete graph contains 17,932 nodes (proteins, complexes, genes, transcription or protein modification events) and 27,976 edges (activation or inhibition). To orient our analysis to the expression profiles and to the biological problem at hand, we built a subgraph with signed edges by extracting the downstream events from three signaling pathways (IL6/IL6-R, IGF1/IGF1-R and CD40), all of which are known to include cellular receptors involved in MM³¹, to the over- and underexpressed variant genes from all datasets by the shortest paths. This cycled, directed subgraph was then filtered by deleting all nodes that are not observed and with one predecessor or one successor³². This filtering step involves no loss of information with respect to the graph coloring model and allowed us to reduce the complexity of the analysis while maintaining the dependencies between the nodes.

Sign consistency modeling framework. In Fig. 1, we illustrate the input (network and transcriptomic data) and output (sign projections) information obtained when the sign consistency modeling is applied. In the following sections, we describe in detail the main modeling steps of this framework.

Graph coloring model. Assuming a directed graph $G(V, E, \alpha)$ in which V is the set of nodes, E is the set of edges and α is a function labeling the edges as $\alpha: E \rightarrow \{+, -\}$, let β be a set of observed data with $\beta: V \rightarrow \{+, -, 0\}$. In our case, β is obtained from GEP and labels only the nodes that are preceded by a “transcription” event as reported by the PID-NCI database. Thus, there are nodes in V , such as proteins or complexes, that remain unlabeled. Our reasoning framework expresses that there is at least one state or *coloring model* of this biological system. A coloring model is an assignment $\mu: V \rightarrow \{+, -, 0\}$ of each node in V to a sign in $\{+, -, 0\}$. Let us denote by S the set of all possible coloring models; note that $|S| = 3^{|V|}$. When imposing the restrictions of β to S , we reduce the size of all possible coloring models (S^*) to $3^{|V|-|\beta|}$.

Sign consistency. The sign consistency imposes a reasoning mode over a graph G and a labeling β (Supplementary Material, Figure S4). This reasoning imposes that each $\{+, -\}$ variation (in a given coloring model) associated with a node n in V is explained by the variations in the direct predecessors of n in G . This notion can be implemented with the following consistency rules, all of which can be verified automatically:

1. All the nodes fixed to be *inputs* are consistent. Usually, these nodes have no predecessors.
2. Each $\{+, -\}$ variation associated with a node n in a given coloring model has to be explained by a direct predecessor of n . That is, each variant node associated with a sign in $\{+, -\}$ that is not an input needs at least one activator (inhibitor) with the same (opposite) sign.
3. Each invariant node m (associated with sign 0) has to be explained either by the fact that (i) all direct predecessors of m are associated with an invariant sign, or (ii) at least two direct predecessors of m are associated with opposite variant signs $\{+, -\}$.

When a graph G is consistent with β , then a set $\bar{S} \subseteq S^*$ of consistent coloring models can be built. A consistent solution in \bar{S} will be a coloring model in which all the nodes of the graph are colored with respect to β and respect the consistency rules.

Repairs. When a consistent solution does not exist, the graph topology of G is not able to explain the labeling β according to the three previously explained consistency rules. In this study, we used two approaches to restore the consistency.

MCOS-repair: This repair mode corrects the graph topology. Considering that the graph is not complete (missing information, generation method, etc.), we can suppose that some inconsistencies are caused by events that are missing from the graph. It is possible to correct the graph by adding a set of artificial influences (Supplementary Material, Figure S5). In this case, we use the *cardinal minimal correction set* (MCOS) of artificial influences that can be added to restore the consistency. The MCOS is in general not unique.

SCENFIT-repair: This repair mode corrects β by considering wrong information in the observed data. β will be corrected by switching the sign of the observed nodes so as to minimize the number of switches. The switch of an observed node is quantified by a cost, as described below. The set of possible minimal SCENFIT-repairs is not unique.

1. Changing a variant sign (+, -) to the opposite variant sign will have a cost of 2.
2. Changing an invariant (respectively variant) into a variant (respectively invariant) sign will have a cost of 1.

Sign projection. After applying a repair operation, the set of consistent coloring models will be the *union* of the consistent coloring models under each minimal repair. Usually, the number of consistent coloring models is very large, and we use a projection of these models to deduce and propose insights from the graph-observations confrontation. We distinguish 7 sign projections classes:

1. 3 classes are *strong*, meaning that the node has the same sign in all consistent solutions $\{+, -, 0\}$.
2. 3 classes are *weak*, meaning that the node has 2 signs in the consistent solutions: *Not+* (-, 0), *Not-* (+, 0), *change* (+, -).
3. The last class means that a node has three signs in the consistent solutions: (+, -, 0).

Key nodes identification. In this analysis, we used the sign projections computed after restoring the consistency using the MCOS-repairs. To compare predictions across individuals using statistical and machine-learning approaches, we decomposed each sign projection result over a node i in V into a triplet of boolean values. The boolean value expresses whether the couple (i, s) , where $s \in \{+, -, 0\}$, belongs to the sign-projection result. Since we only focused on sign-projections, the nodes observed in β were not considered. To reduce the number of variables, we excluded the boolean value that refers to invariant couples (nodes coupled with “0”). In this way, we represent the sign-projections obtained for each GEP as a boolean matrix M of size $2 \times m \times (N^{MC} + N^{NPC})$, where m represents the number of nodes in G that were never observed in any GEP and N^{MC} (respectively N^{NPC}) represents the number of profiles in class MC (respectively NPC). M_{ij} stands for the decomposed prediction of node i under profile j ; note that M_{ij} can be separated into M_{ij}^+ and M_{ij}^- , where $M_{ij}^s = 1$ expresses that node i is predicted to be of sign s in profile j . To identify specific markers of MM, we analyzed M and looked for overrepresented values when comparing the vectors belonging to MC with those belonging to NPC. For this, we used two approaches, a machine-learning approach based on supervised learning and a

statistical approach based on frequency classification. For the supervised learning, we used a decision tree³³ and a random forest classification³⁴. Due to the underrepresentation of the NPC, we increased the weight of each NPC by 67 so as to have the same order of population in each group (9 NPC and 602 MC). For the frequency approach, we calculated the frequency score (FS) for each group (MC or NPC) and for each assignment (i, s) as follows:

$$FS_{i,s}^C = \frac{1}{N^C} \sum_{j=1}^{N^C} M_{ij}^s, \quad (1)$$

where C represents the class MC or NPC and s represents the $\{+, -\}$ sign assigned to i . We then sorted our results based on a Fisher test between the proportions for NPC and MC to determine the most specific node assignments for the MC datasets.

Nodes perturbation. In this analysis, we quantified the effectiveness of a node perturbation (Supplementary Material, Figure S6) to simulate *in silico* the activation or inhibition of a protein. The quantification of these *in silico* perturbations was performed in two steps. First, we considered the set of assignments in M (see previous section), where $M_{ij}^s = 1$ expresses that node i is predicted to be of sign s in profile j , with $s \in \{+, -\}$. For each assignment M_{ij}^s , we generated a new dataset of observations β_{ij}^s identical to the original dataset of profile j (β_j) except that we added an observation on node i fixed to $s \in \{+, -\}$. We then computed the SCENFIT score SF_{ij}^s between the graph G and β_{ij}^s . The second step consisted of computing the Top Perturbation Score (TPS) for each assignment (i, s) according to its SF_{ij}^s across all GEP j , as follows:

$$TPS_{i,s}^C = \frac{1}{N^C} \sum_{j=1}^{N^C} f(i, s, j),$$

where

$$f(i, s, j) = \begin{cases} 1, & \text{if } SF_{ij}^s \geq \text{top}(SF_{kj}^s), \forall k \in V \setminus \text{Dom}(\beta_{ij}^s). \\ 0, & \text{otherwise.} \end{cases}$$

In these equations, C represents the class MC or NPC, and s represents the $\{+, -\}$ sign assigned to i . The function $\text{top}(SF_{kj}^s)$ will compute the threshold score that separates the 10% top-ranked SCENFIT scores of profile j , that is, those perturbations that generate the highest number of SCENFIT repairs.

Software and tools. For the sign consistency analysis, we used IGGY²⁹, which makes use of an ASP³⁵ description of the consistency problem. The graph generation and the mapping of predictions to the couples node-sign were implemented with Python 2.7 using the package NetworkX³⁶. The learning and statistical analysis was conducted using R³⁷. The computation associated with testing the consistency of the 611 GEP required 5 minutes on a standard machine. All the calculations of nodes perturbations were conducted using the BIRD infrastructure (www.pf-bird.univ-nantes.fr) with 320 nodes and 1.3To RAM.

Graphs availability. All graphs used in this study are available online using cynetshare. The subgraphs of NCI-PID before (goo.gl/upfzwc) and after compaction (goo.gl/SfNSv4). The subgraph from Fig. 2 is available at goo.gl/YgHvtQ. The cytoscape session containing all graphs is available at goo.gl/V1Rno5.

Results

Data discretization and graph generation. The NCI-PID integration allowed us to find 634 genes (a protein preceded by a transcription event). Independently, our discretization method (Supplementary Material, Figure S1) proposed observations $\{+, -, 0\}$ on microarray probes corresponding to 15,418 proteins identified in Uniprot. Merging both lists allowed us to identify 557 genes present in the NCI-PID and observed as over/underexpressed or invariant in our GEPs. Variant and invariant genes are distributed across the MC and NPC datasets (Table 1). By extracting the downstream events from three signaling pathways (IL6/IL6-R, IGF1/IGF1-R and CD40)³¹ to the variant genes, we generated an induced subgraph from NCI-PID containing 2,269 nodes, 2,683 edges and connecting 529 variant genes. This graph was then compacted to a new graph with 596 nodes and 960 edges (Fig. 2) and composed of 529 observed nodes (genes) and 67 unobserved nodes, including 23 proteins, 33 complexes, 2 biological processes, 9 proteins reactions (translocation, phosphorylation, etc.).

Validation of predictions. The confrontation between the data and the graph topology allowed us to predict the node signs for each dataset (Table 1). To validate our predictions, we compared the precision of the predictions with that of the randomized data. In this case, we used 50% of the measured genes $\{+, -, 0\}$ to predict the other half of the genes for each sample; we performed the same process after randomizing the data and repeated this computation up to 1000 times. We obtained two sets of precisions (Fig. 3; Supplementary Material, Table S1). A two-tailed t-test yielded a p-value lower than $2.2e-16$. This shows the efficiency of our prediction method in comparison with random precision.

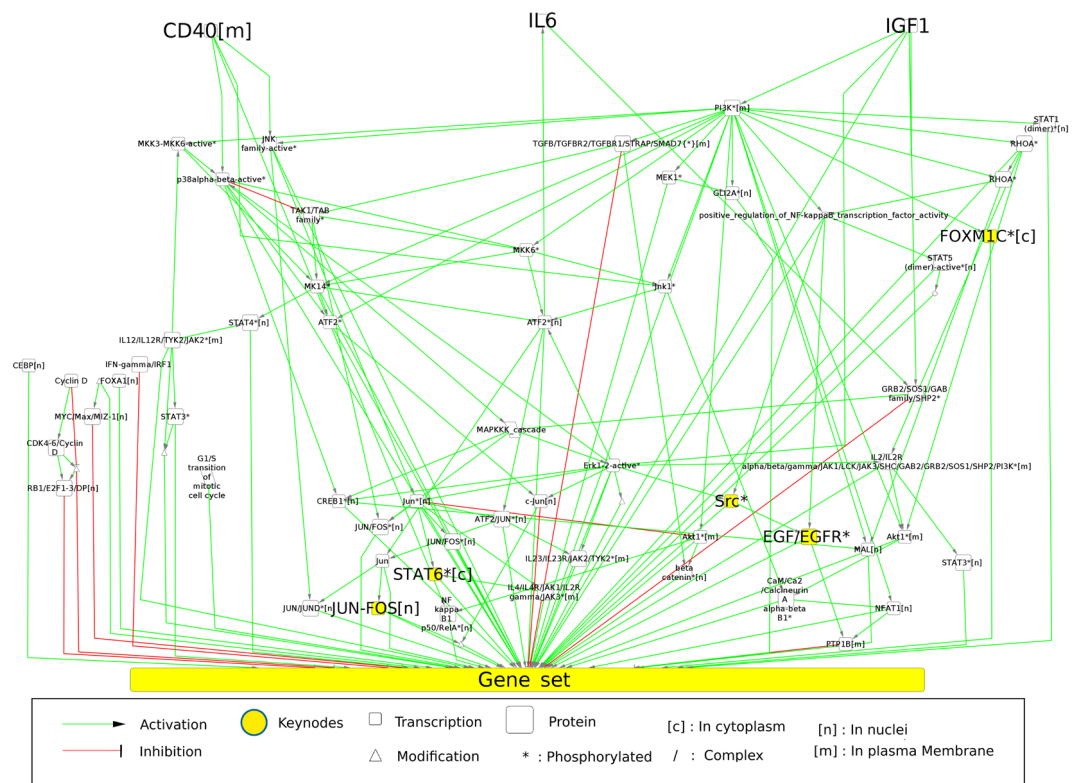


Figure 2. Representation of the subgraph obtained from the PID-NCI database. CD40, IL6 and IGF1 (the nodes in the top portion of the graph) are the 3 queried pathways. The 529 genes that are differentially expressed across all profiles are merged for this representation in the node “Gene set”. The 5 top-ranked nodes according to their FS are labeled in bold type and colored in yellow. We used the same syntax for all nodes in this study. The edges from the “Gene set” node to proteins have been deleted for the sake of clarity.

Signs	Observed data		Predicted data	
	NPC	MC	NPC	MC
+	34%	38%	30%	31%
-	34%	51%	29%	36%
0	32%	11%	14%	3%
change	—	—	7%	6%
Not+	—	—	2%	1%
Not-	—	—	3%	1%
?	—	—	15%	22%
Total	2085	210975	3279	153181

Table 1. Observed and predicted data repartition between NPC and MC. Observed data are the data extracted from the gene expression profiles. Predicted data are the sign projections predicted after the confrontation between the observed data and the PID-NCI graph. In the last row, we show the total observations and predictions across all profiles.

Identification of specific node assignments for MC. To identify MC subgroups, we applied a decision tree algorithm to the presence/absence value of a sign prediction (see Methods section). This result is illustrated in Fig. 4. It shows that the combination of the assignments (JUN/FOS[n, -) and (FOXM1*[c, -) is associated with the majority of MC (73%) and that the method can distinguish MC from NPC. JUN/FOS[n] represents the protein complex composed of JUN and FOS, which is located in the nucleus, whereas FOXM1*[c] represents the FOXM1 protein, which is phosphorylated and located in the cytoplasm. The full node syntax is given in Fig. 2. Moreover, we can identify another important group of MC (13%) that is characterized by the presence of (JUN/FOS[n, -) and the absence of (FOXM1*[c, -) and (SRC*, -). Similar results were obtained using a random forest classification (Supplementary Material, Figure S7).

To characterize the shared specificity for all MC, we computed the frequency scores (FS) for our predictions (see Methods section). The complete list of the FS obtained is shown in Table S2 of the Supplementary Material. In Table 2, we show the 5 best p-values associated with a Fisher test with $FS_{MC} > FS_{NPC}$. For these

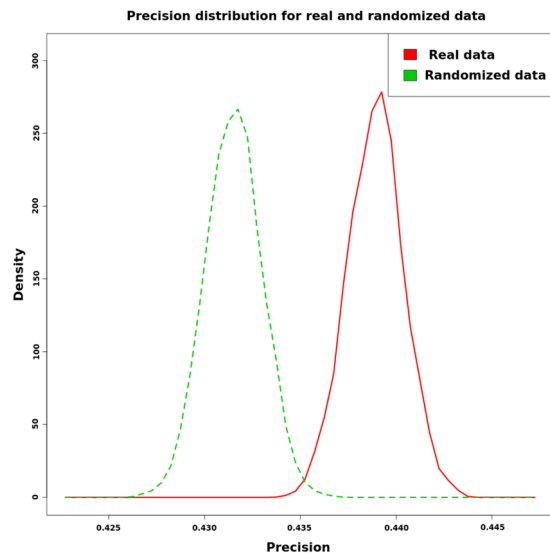


Figure 3. Precision distribution of our method with real observed and randomized data.

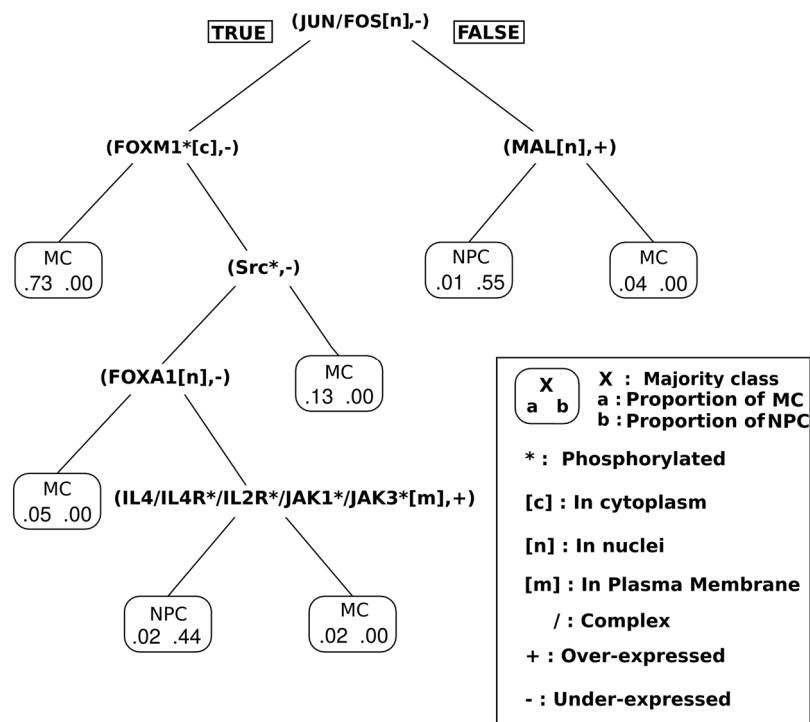


Figure 4. Decision tree based on predicted node-sign assignments.

assignments, we checked the number of input/output variant genes connected to each node in the graph (Table 2, column connectivity). We observe that inhibition of the complex JUN/FOS[n] is predicted for 95.6% of MC. The activity levels of FOXM1*[c] and STAT6*[c] were predicted to decrease. This decrease, in terms of protein activity, is correlated with the level of gene expression in 76% and 93%, respectively, of the MC datasets. The FS classification identified the presence of (Src*, +) as an interesting marker for MC datasets. Interestingly, the decision tree approach identified the absence of (Src*, -) as distinguish MC datasets that were previously characterized by (JUN/FOS[n], -) and (FOXM1*[c], -). Both the machine-learning and statistical methods identified (JUN/FOS[n], -) and (FOXM1*[c], -) as important markers of MC datasets. In Fig. 2, we show (marked as yellow nodes) how these 5 main proteins or protein complexes appear connected following the PID-NCI representation.

Predicted node	Sign	FS^{NPC}	FS^{MC}	p.val (Fisher)	References	Connectivity	OVE	
							+	-
JUN/FOS[n]	-	0.444	0.956	2.65E-005	38–42	8/529	373	137
FOXM1*[c]	-	0.222	0.774	7.97E-004	43, 44	529/529	85	265
STAT6*[c]	-	0.222	0.764	1.05E-003	∅	8/529	30	429
EGF/EGFR*[m]	+	0.556	0.935	2.08E-003	45–47	529/529	79	4
Src*	+	0.556	0.935	2.08E-003	48–50	529/529	110	48

Table 2. 5 top-ranked results for the frequency analysis for MC signatures. FS^{NPC} and FS^{MC} show the frequency scores for NPC and MC, respectively. The references column lists the publications that agreed with our sign prediction. Connectivity refers to the ratio of genes connected to each predicted node. The OVE (observed variant expression) shows the repartition of variant gene expression using the best precision threshold without considering graph information.

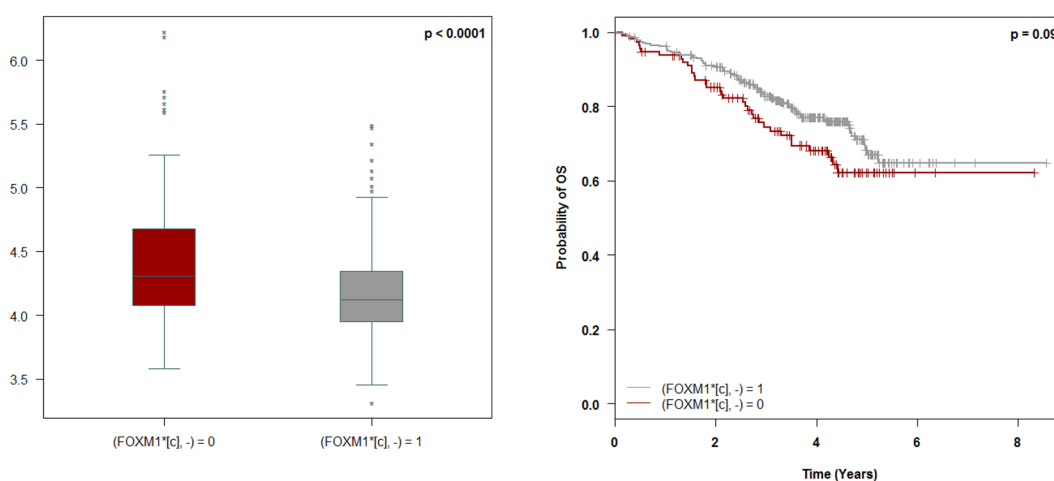


Figure 5. (Left) Gene expression of FOXM1 among MC datasets with or without the prediction (FOXM1*[c], -). (Right) Overall survival (OS) of patients with prediction (FOXM1*[c], -) or without prediction (FOXM1*[c], -).

JUN/FOS activity as specific marker. The FOS and JUN proteins form a heterodimer complex that is responsible for AP-1 activity. This activity is known to play a role in tumorigenesis because it has been implicated in the induction of apoptosis, in the promotion of cell survival and in proliferation. The classification methods showed that (JUN/FOS[n], -) is the best assignment to distinguish MC from NPC and revealed that AP-1 activity is lower in almost all MM patients than in normal controls. Inspection of individual patients' subgraphs showed predominantly underexpression (65% of the observed expression in MC) of the proapoptotic protein BIM (Supplementary Material, Figure S8). These results are in agreement with the results of *in vitro* studies demonstrating that in myeloma cell lines IL6 protects against apoptosis via AP-1 inactivation³⁹.

FOXM1 activity as survival marker. FOXM1, a transcriptional factor known to be associated with MM, has been studied as a therapeutic target⁴⁴. Based on the graph reduction and on our reasoning model, FOXM1*[c] is equivalent to FOXM1*[n] and is representative of the FOXM1 transcriptional activity. Firstly, we analyzed FOXM1 gene expression in the MC groups in which FOXM1 activity was predicted. We found that decreased FOXM1 activity is associated with reduced expression of the FOXM1 gene (Fig. 5, left). Since our model identified a subgroup of patients with decreased activity of FOXM1 and since decreased expression of FOXM1 is associated with superior survival (Supplementary Material, Figure S9), we wanted to know whether FOXM1 activity could impact survival. We compared overall survival (OS) in both predicted groups in the larger cohort of patients that received comparable treatment (Velcade-dexamethasone induction followed by high-dose melphalan and autologous stem cell transplantation; $n = 450$) (Fig. 5, right). A log-rank test between these groups yielded a p-value of < 0.1 , allowing us to conclude that low FOXM1 activity is associated with a trend towards better survival.

Improvement of the current prognostic model in MM using node variables. Univariate and multivariate Cox proportional hazards analyses were performed on the cohort of 450 MM patients who received comparable treatment to determine the relative prognostic values of the 201 couples combining unobserved nodes and all signs (+, -, 0) and the three strongest known prognostic variables in MM (Table 3); these were the translocation of chromosomes 4 and 14 (t(4;14)), the deletion in the short arm of chromosome 17 (del(17p)) and serum 2-microglobulin ≥ 5.5 mg/L (β_2 -microglobulin) for OS determination⁵¹. In the multivariate analysis,

Parameters	Univariate analysis			Multivariate analysis		
	HR	95%CI	Pvalue	HR	95%CI	Pvalue
β_2 -microglobulin, mg/L ≥ 5.5 v < 5.5	2.03	1.35–3.05	0.001	1.53	0.99–2.35	0.056
t(4,14), Yes v no	3.19	2.08–4.89	<0.01	2.41	1.49–3.90	<0.01
del17p > 60 v ≤ 60	4.16	2.53–6.83	<0.01	3.16	1.80–5.56	<0.01
(G1/S transition of mitotic cell cycle, –), yes v no	0.33	0.22–0.47	<0.01	0.47	0.30–0.72	<0.01
(RB1/E2F1–3/DP[n], +), yes v no	0.49	0.33–0.75	0.001	0.58	0.36–0.93	0.025

Table 3. Parameters Associated With Overall Survival.

Node	Dir.	TPS^{NPC}	TPS^{MC}	p.val
JUN/FOS[n]	+	22.2%	74.6%	0.001
	–	44.4%	0.5%	1
FOXMI*[c]	+	11.1%	36.4%	0.107
	–	55.6%	19.1%	0.997
STAT6*[c]	+	33.3%	55.0%	0.169
	–	44.4%	21.9%	0.970
EGF/EGFR*[m]	+	0.0%	0.3%	0.971
	–	0.0%	3.5%	0.728
Src*	+	0.0%	1.3%	0.887
	–	11.1%	33.4%	0.150

Table 4. Top perturbation score for nodes identified with the FS method. Dir stands for the direction of the perturbation (+, activation and –, inhibition). TPS represents the frequency with which perturbing a node in a specific direction was significant (i.e. it generated a high, 10% top, SCENFIT score) across the MC profiles (TPS^{MC}) or NPC profiles (TPS^{NPC}). The bold percentages refer to perturbations that have a direction opposite to that of the predicted signs obtained with the frequency score (Table 3). Pval was obtained using a unilateral Fisher test.

estimation of hazard ratios for death indicates that both (G1/S transition of mitotic cell cycle, –) and (RB1/E2F1-3/DP[n], +) were independent powerful prognostic factors (Supplementary Material, Figure S10).

The multivariate model with the known prognostic parameters shows that these factors increase the log-likelihood from –515.16 (null model) to –496.62 (3 parameter model), with p-significance $< 10^{-7}$ (null model vs 3 parameter model) whereas the parameters (G1/S transition of mitotic cell cycle, –) and (RB1/E2F1-3/DP[n], +) increase the log-likelihood from –496.62 (3-parameter model: $AIC3p = 999.2$) to –486.90 (5-parameter model: $AIC5p = 983.8$) with p-significance $< 10^{-4}$ (3-parameter model vs 5-parameter model) and $< 10^{-10}$ (null model vs 5-parameter model). Therefore, we can conclude that the 5-parameter model provides more prognostic information than the 3-parameter model ($AIC5p < AIC3p$ and $p5p$ vs $3p < 10^{-4}$). In term of the global increase in the log-likelihood between the null model and the 5-parameter model, the specific impact of the selected pairs represents more than 34% of the total.

Node perturbation. From the computation of all *in silico* node perturbations (see Methods section), we evaluated the impact of perturbing the key nodes found with the FS method (Table 4). A unilateral Fisher test allowed us to evaluate the significance of each perturbation compared to the NPC datasets. We can see that the activation of JUN/FOS generates a top-ranked (10% top) score of conflicts and therefore repairs 74.6% of the MC datasets, whereas it repairs only 22.2% of the NPC datasets. Interestingly, *in vitro* JUN overexpression in MM cell lines results in cell death and growth inhibition⁴¹. A similar tendency (more conflicts in MC than in NPC) is observed when FOXMI is activated, but the difference cannot be considered significant. Nonetheless, we note that of the 36.4% of profiles in which the activation of FOXMI is top-ranked, 96.8% correspond to patient profiles with the prediction (FOXMI*[c], –) (Supplementary Material, Table S3). For the other proteins and complexes, we can see that the difference between MM and NPC is not significant. It is worth noting that the p-value of a perturbation that goes in the opposite direction of the prediction shown in Table 2 is in all cases lower than the one of a perturbation which goes in the same direction of the prediction.

Discussion

Data discretization and graph generation. Our method incorporates both differential and similar expressions in its reasoning. All *pathway analysis* methods reviewed in the Introduction use the difference in gene expression between the two classes of subjects to extract the specific signatures. The similarity of expression between classes is not used in the ORA and FCS approaches because these methods base their analyses on the differential expression of genes. The PT approaches reviewed use only differential gene expression in their reasoning. We believe that adding information on similar expression enables us to better capture cellular behavior.

The results of the precision analysis using a two-sign coloring model tend to support this hypothesis. Our method differs from classic pathway analysis methods in that it incorporates the notion of automatic reasoning. Within the context of MM, we are able to automatically detect repairs. These repairs are specific for each GEP and could represent cancer mutations, regulatory network incompleteness or experimental errors. The graph used in this study contains 529 genes; we can therefore observe the strong connectivity that exists among PID pathways since the total number of genes in the PID is 634. This strong connectivity is important for methods such as ours that are able to reason on the information content of the whole database. We observe, however, that the number of genes connected to cancer pathways in PID is far below the total number of human genes. This underrepresentation of regulatory knowledge is an important limitation of PID. Apart from this fact, PID-NCI includes important modeling information that identifies *transcription events*. This information allows us to separate gene expression from protein activity. These two parameters are not necessarily correlated, especially in cases involving phosphorylated proteins or complexes such as JUN/FOS.

Key nodes identification. Our analysis of the predictions made by the method allowed us to identify nodes associated with a sign specific to MC compared to NPC datasets. Among these assignments, we found the inhibition of JUN/FOS[n] and FOXM1*[c]. These proteins are known to be involved in cancer in general^{52,53} and in hematological malignancies in particular^{43,44}. In the case of FOXM1, we showed that this transcription factor can represent a survival marker when its activity decreases. We can draw a parallel with the bibliography, which identifies the activation of the FOXM1 pathway as a risk factor. For JUN/FOS, our analysis identified this pathway as a potential therapeutic target but not as a survival marker. We observed that inhibition of the associated pathways has been already identified in MM patients^{39,41} and in patients with other cancers. Moreover, this pathway is targeted in some therapeutic approaches^{38,40}. We identified two couples that improve classical prognostic models. In the case of the first couple, (G1/S transition of the mitotic cell cycle, -), we can associate this node with the proliferation pathway. The computed prognostic model showed that the prediction of inhibited proliferation can be a protective factor for MM patients. The second node, (RB1/E2F1-3/DP[n], +), was also identified as a protective factor by the 5-parameter model. This complex is known to be involved in the RB pathway, which influences cell growth pathways by regulating the initiation of DNA replication. This pathway is usually altered in cancer, leading to a loss of function⁵⁴, and current therapeutic approaches aim to activate this pathway⁵⁵.

Using the *in silico* node perturbation method, we were able to estimate the effect of perturbing a node within a particular dataset (*i.e.* single patient cancer cell). This method represents a powerful tool for analyzing the consequences of perturbations of oncogenic pathways in a given patient, especially as *in-vitro* experiments are limited due to the small amount of viable myeloma cells that are obtained after bone marrow aspiration. The results of this *in silico* analysis show that activation of JUN/FOS[n] had a significant impact on 75% of MC profiles; all of these JUN/FOS[n] = "+" sensible MC profiles had the prediction (JUN/FOS[n], -). In addition, activating FOXM1*[c] had a significant impact on 36.4% of the profiles; 96.8% of the FOXM1*[c] = "+" sensible MC profiles had the prediction (FOXM1*[c], -). The difference in the percentages of JUN/FOS[n] and FOXM1*[c] can be explained by the graph topology and the connectivity of the individual nodes. JUN/FOS[n] is connected to eight genes through a distance of 1 molecular species (Supplementary Material, Figure S8); therefore, perturbing JUN/FOS[n] will impact these genes directly since they are strongly constrained by the sign of JUN/FOS[n]. On the other hand, FOXM1 is connected to 529 genes through longer paths through distances of from 4 to 77 molecular species. These genes may have other predecessors that are independent of FOXM1; this could explain why activation of FOXM1 has a strong effect on only 37% of the MC profiles. Overall, we think that this *in silico* method could be used to reinforce the choice of a therapeutic target for a specific patient profile.

Conclusion

In this study, we used a specific approach to study and understand the heterogeneous gene expression profiles of approximately 600 multiple myeloma (MM) patients. Our primary goal was to provide mechanistic scenarios by identifying protein activity states of molecules that may be central to the diversity of gene expression. Our approach relies heavily on reasoning based on graphs and on changes in gene expression in the form of logical programs that combine these two types of information. The method proposed here can be summarized in the following steps. First, we obtained a directed graph, allowing us to connect significantly up-/down-regulated genes to upstream MM-related cellular receptors. Second, we confronted this graph to transcriptomic data with IGGY, which is a tool that reasons on the logic of the graph and on shifts of expression in the data so as to predict (*node, sign*) assignments representing the specific states of biological entities. Using two approaches of classification, we were able to identify specific assignments for MC datasets compared to NPC datasets. Finally, taking advantage of our modeling framework, we studied the effect of performing single *in silico* perturbations.

One advantage of this method is that it makes it possible to infer information about protein states from transcriptomic data by using the causal nature of the interactions as documented in PID. This can be interesting when constructing biological models and, more specifically, when developing cancer models for which proteomic data are not always available and extractable, whereas transcriptomic data are easier to obtain. Moreover, compared to the previously presented classical pathway analysis methods, we identify not only the specific biological processes that are implicated in cancer profiles but also the mechanisms associated with those phenomena. After statistically testing the quality of the method's predictions, we proposed a set of five top-scoring proteins based on their respective changes in activity in MC compared with NPC. We found the AP-1 complex and the FOXM1 transcription factor to be concomitantly inactivated in a strong majority of patients regardless of treatment or age. Interestingly, this method identified a subgroup of MM patients with increased FOXM1 activity associated with poor survival. These findings allow us to validate the predictions of our approach and show that it is feasible to individualize or restrict the analysis of multiple expression profiles to identify markers within subgroups of

profiles and to identify parameters associated with survival in these subgroups. The 5-parameter model including the two predicted nodes improves the standard prognostic model in MM. In addition to its strong prognostic value, our model revealed two nodes, (G1/S transition of mitotic cell cycle, $-$) and (RB1/E2F1-3/DP[n], $+$), that are of potential biological interest in the understanding of the molecular mechanisms underlying resistance to treatment. Note that these nodes can only be predicted with the graph and coloring model, since they are a (logical) consequence of the GEP. Our results on *in silico* perturbations of a system are also encouraging because they show that changes in the activity of the predicted proteins can serve as input information for conducting efficient perturbations. In this work, we focused only on single perturbations, since they are more experimentally realistic. As a perspective of this work, we wish to deepen the graph vs. gene-expression confrontation analysis so as to understand the differences between MM subgroups based on age, prognosis and other criteria. In this context, one line of research would be to study minimal subsets of perturbations. Another possible line of research would be the classification of gene expression profiles based on plausible graph-coloring models.

References

- Morgan, G. J., Walker, B. A. & Davies, F. E. The genetic architecture of multiple myeloma. *Nature reviews. Cancer* **12**, 335–48 (2012).
- Zhan, F. *et al.* The molecular classification of multiple myeloma. *Blood* **108**, 2020–2028 (2006).
- Decaux, O. *et al.* Prediction of survival in multiple myeloma based on gene expression profiles reveals cell cycle and chromosomal instability signatures in high-risk patients and hyperdiploid signatures in low-risk patients: A study of the Intergroupe Francophone du Myélome. *Journal of Clinical Oncology* **26**, 4798–4805 (2008).
- Shaughnessy, J. D. *et al.* A validated gene expression model of high-risk multiple myeloma is defined by deregulated expression of genes mapping to chromosome 1. *Blood* **109**, 2276–84 (2007).
- Avet-Loiseau, H. *et al.* Prognostic significance of copy-number alterations in multiple myeloma. *Journal of clinical oncology: official journal of the American Society of Clinical Oncology* **27**, 4585–90 (2009).
- Broyl, A. *et al.* Gene expression profiling for molecular classification of multiple myeloma in newly diagnosed patients. *Blood* **116**, 2543–53 (2010).
- Walker, B. A. *et al.* Mutational Spectrum, Copy Number Changes, and Outcome: Results of a Sequencing Study of Patients With Newly Diagnosed Myeloma. *Journal of clinical oncology: official journal of the American Society of Clinical Oncology* **33**, 3911–20 (2015).
- Rashid, N. U. *et al.* Differential and limited expression of mutant alleles in multiple myeloma. *Blood* **124**, 3110–7 (2014).
- Mansour, M. R. *et al.* Oncogene regulation. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science (New York, NY)* **346**, 1373–7 (2014).
- The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Nature Genetics* **25**, 25–29 (2000).
- Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research* **28**, 27–30 (2000).
- Schaefer, C. F. *et al.* PID: the Pathway Interaction Database. *Nucleic acids research* **37**, D674–9 (2009).
- Kelder, T. *et al.* WikiPathways: building research communities on biological pathways. *Nucleic acids research* **40**, D1301–7 (2012).
- Wingender, E. The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. *Briefings in Bioinformatics* **9**, 326–332 (2008).
- Boué, S. *et al.* Causal biological network database: a comprehensive platform of causal biological network models focused on the pulmonary and vascular systems. *Database: the journal of biological databases and curation* **2015**, bav030 (2015).
- Khatri, P., Sirota, M. & Butte, A. J. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS computational biology* **8**, e1002375 (2012).
- Bindea, G. *et al.* ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics (Oxford, England)* **25**, 1091–3 (2009).
- Catlett, N. L. *et al.* Reverse causal reasoning: applying qualitative causal knowledge to the interpretation of high-throughput data. *BMC bioinformatics* **14**, 340 (2013).
- Martin, F. *et al.* Quantification of biological network perturbations for mechanistic insight and diagnostics using two-layer causal models. *BMC bioinformatics* **15**, 238 (2014).
- Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 15545–50 (2005).
- Backes, C. *et al.* GeneTrail—advanced gene set enrichment analysis. *Nucleic acids research* **35**, W186–92 (2007).
- Lefebvre, C. *et al.* A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers. *Molecular systems biology* **6**, 377 (2010).
- Kong, S. W., Pu, W. T. & Park, P. J. A multivariate approach for integrating genome-wide expression data and biological knowledge. *Bioinformatics (Oxford, England)* **22**, 2373–80 (2006).
- Ideker, T., Ozier, O., Schwikowski, B. & Siegel, A. F. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics (Oxford, England)* **18**(Suppl 1), S233–40 (2002).
- Komurov, K., Dursun, S., Erdin, S. & Ram, P. T. NetWalker: a contextual network analysis tool for functional genomics. *BMC genomics* **13**, 282 (2012).
- Liu, W. *et al.* Topologically inferring risk-active pathways toward precise cancer classification by directed random walk. *Bioinformatics (Oxford, England)* **29**, 2169–77 (2013).
- Yaveroğlu, Ö. N., Milenković, T. & Pržulj, N. Proper evaluation of alignment-free network comparison methods. *Bioinformatics* **31**, 2697–2704 (2015).
- Draghici, S. *et al.* A systems biology approach for pathway level analysis. *Genome research* **17**, 1537–45 (2007).
- S, T. *et al.* Extended notions of sign consistency to relate experimental data to signaling and regulatory network topologies. *BMC Bioinformatics* **16**, 345 (2015).
- Avet-Loiseau, H. *et al.* Genetic abnormalities and survival in multiple myeloma: the experience of the Intergroupe Francophone du Myélome. *Blood* **109**, 3489–95 (2007).
- Klein, B. Positioning NK- κ B in multiple myeloma. *Blood* **115**, 3422–4 (2010).
- Saez-Rodriguez, J. *et al.* Discrete logic modelling as a means to link protein signalling networks with functional analysis of mammalian signal transduction. *Molecular systems biology* **5**, 331 (2009).
- Quinlan, J. Simplifying decision trees. *International Journal of Man-Machine Studies* **27**, 221–234 (1987).
- Breiman, L. Random Forests. *Machine Learning* **45**, 5–32 (2001).
- Baral, C. *Knowledge Representation, Reasoning and Declarative Problem Solving* (Cambridge University Press, 2003).
- Hagberg, A. A., Schult, D. A. & Swart, P. J. Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference (SciPy2008)*, 11–15 (2008).
- R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2015).

38. Podar, K. *et al.* Up-regulation of c-Jun inhibits proliferation and induces apoptosis via caspase-triggered c-Abl cleavage in human multiple myeloma. *Cancer research* **67**, 1680–8 (2007).
39. Xu, F. H. *et al.* Interleukin-6-induced inhibition of multiple myeloma cell apoptosis: support for the hypothesis that protection is mediated via inhibition of the JNK/SAPK pathway. *Blood* **92**, 241–251 (1998).
40. Saha, M. N. *et al.* Targeting p53 via JNK pathway: a novel role of RITA for apoptotic signaling in multiple myeloma. *PLoS one* **7**, e30215 (2012).
41. Chen, L. *et al.* Identification of early growth response protein 1 (EGR-1) as a novel target for JUN-induced apoptosis in multiple myeloma. *Blood* **115**, 61–70 (2010).
42. Fan, F. *et al.* Targeting Mcl-1 for multiple myeloma (MM) therapy: drug-induced generation of Mcl-1 fragment Mcl-1(128-350) triggers MM cell death via c-Jun upregulation. *Cancer letters* **343**, 286–94 (2014).
43. Uddin, S. *et al.* Overexpression of FoxM1 offers a promising therapeutic target in diffuse large B-cell lymphoma. *Haematologica* **97**, 1092–100 (2012).
44. Gu, C. *et al.* FOXM1 is a therapeutic target for high-risk multiple myeloma. *Leukemia* **30**, 873–882 (2016).
45. Mahtouk, K. *et al.* An inhibitor of the EGF receptor family blocks myeloma cell growth factor activity of HB-EGF and potentiates dexamethasone or anti-IL-6 antibody-induced apoptosis. *Blood* **103**, 1829–37 (2004).
46. Mahtouk, K. *et al.* Expression of EGF-family receptors and amphiregulin in multiple myeloma. Amphiregulin is a growth factor for myeloma cells. *Oncogene* **24**, 3512–3524 (2005).
47. Johnston, J. B. *et al.* Targeting the EGFR pathway for cancer therapy. *Current medicinal chemistry* **13**, 3483–3492 (2006).
48. Hallek, M. *et al.* Signal transduction of interleukin-6 involves tyrosine phosphorylation of multiple cytosolic proteins and activation of Src-family kinases Fyn, Hck, and Lyn in multiple myeloma cell lines. *Experimental hematology* **25**, 1367–77 (1997).
49. Coluccia, A. M. L. *et al.* Validation of PDGFRbeta and c-Src tyrosine kinases as tumor/vessel targets in patients with multiple myeloma: preclinical efficacy of the novel, orally available inhibitor dasatinib. *Blood* **112**, 1346–56 (2008).
50. Ishikawa, H. Requirements of src family kinase activity associated with CD45 for myeloma cell proliferation by interleukin-6. *Blood* **99**, 2172–2178 (2002).
51. Avet-Loiseau, H. *et al.* Combining fluorescent *in situ* hybridization data with ISS staging improves risk assessment in myeloma: an International Myeloma Working Group collaborative project. *Leukemia* **27**, 711–717 (2013).
52. Eferl, R. & Wagner, E. F. AP-1: a double-edged sword in tumorigenesis. *Nature reviews. Cancer* **3**, 859–68 (2003).
53. Shaulian, E. & Karin, M. AP-1 as a regulator of cell life and death. *Nature Cell Biology* **4**, E131–E136 (2002).
54. Nevins, J. R. The Rb/E2F pathway and cancer. *Human molecular genetics* **10**, 699–703 (2001).
55. Knudsen, E. S. & Wang, J. Y. J. Targeting the RB-pathway in cancer therapy. *Clinical cancer research: an official journal of the American Association for Cancer Research* **16**, 1094–9 (2010).

Acknowledgements

This study was supported by Intergroupe Francophone du Myélome and by a French Institute National du Cancer Grant EVACAMM PROG/09/10 (to H.A.L., S.M.), a National Institutes of Health Grant PO1CA155258-01 (to S.M., H.A.L., N.C.M.), and a research grant from Celgene. B.M.'s PhD scholarship was funded by GRIOTE project. We would like to thank Elise Douillard, Magali Devic, Emilie Morenton and Nathalie Roi for excellent technical assistance. We thank Jérémie Bourdon, Nathalie Theret and Sophia Tsoka for suggestions and critical reading of the manuscript. We are most grateful to the bioinformatics core facility of Nantes (BiRD - Biogenouest) for its technical support.

Author Contributions

B.M. implemented the predictions analysis and the perturbations methods, performed the computational analysis and wrote the paper. B.M., S.M., F.M. and C.G. conceived and supervised the study and drafted the manuscript. B.M., S.M., O.R., F.M. and C.G. discussed the results of the data analysis. P.D. performed the k-mean discretization method and its comparison and implemented the predictions validation method. W.G. performed the microarray analysis. L.C. performed the survival models comparison with and without predictions. C.G.C. performed the statistical analysis. H.A.L., M.A., T.F., N.C.M., and P.M. provided samples and clinical data. All authors read and approved the final manuscript.

Additional Information

Supplementary information accompanies this paper at doi:10.1038/s41598-017-09378-9

Competing Interests: The authors declare that they have no competing interests.

Accession codes Minimum Information About a Microarray Experiment-compliant data has been deposited at: Gene Expression Omnibus with accession number GSE83503.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2017

.3 Modèle de perturbations multiples

Le modèle de perturbations multiples permet à partir d'un graphe (ligne 1 à 7) sa coloration (ligne 9 à 15) une liste de perturbations possibles (ligne 17 à 21) et enfin un motif à obtenir (ligne 23 à 25), d'identifier la combinaison minimale de perturbations à appliquer pour amener à ce motif pré-déterminé..

```

1 edge (a, c, -1) .
2 edge (c, f, 1) .
3 edge (a, b, 1) .
4 edge (c, g1, -1) .
5 edge (f, g2, 1) .
6 edge (b, g2, 1) .
7 edge (b, g3, 1) .
8
9 color (a, 1) .
10 color (b, 1) .
11 color (c, -1) .
12 color (f, -1) .
13 color (g1, 1) .
14 color (g2, -1) .
15 color (g3, 1) .
16
17 perturbateur (b, 1) .
18 perturbateur (b, -1) .
19 perturbateur (c, 1) .
20 perturbateur (c, -1) .
21 perturbateur (a, -1) .
22
23 patternCible (g1, -1) .
24 patternCible (g2, 1) .
25 patternCible (g3, 1) .
26
27 node (X) ← edge (X, _, _) .
28 node (X) ← edge (_, X, _) .
29
30 sign (1; -1) .
31
32 l{cible (X, Y) : perturbateur (X, Y) } .
33 ← cible (X, Y), cible (X, Z), Y!=Z, sign (Y), sign (Z) .
34
35 perturbe (X) ← cible (X, _) .
36 perturbe (X) ← edge (Y, X, _), perturbe (Y) .
37
38 newColor (X, Y) ← cible (X, Y) .
39 newColor (X, Y*Z) ← newColor (ArcSup, Y), edge (ArcSup, X, Z) .
40 newColor (X, Y*Z) ← perturbe (X), color (ArcSup, Y), edge (ArcSup, X, Z), not cible (X, _), not
    perturbe (ArcSup) .

```



```
41
42 reussi(X) ← newColor(X,Y), patternCible(X,Y), not newColor(X,Z), Z!=Y, sign(Z).
43 reussi(X) ← color(X,Y), patternCible(X,Y), not newColor(X,Z), Z!=Y, sign(Z).
44 sumReussite(X) ← X =#count{ node(Z) :reussi(Z) }.
45 #maximize {X@2 : sumReussite(X)}.
46
47 sumPerturbateurs(X) ← X =#count{ node(Z) :cible(Z,_) }.
48 #minimize {X@1 : sumPerturbateurs(X)}.
49 #show cible/2.
```


Thèse de Doctorat

Bertrand MIANNAY

Analyse de réseaux de régulation par approches de coloration de graphes dans le cadre du myélome multiple

Regulatory networks analysis with graph coloring approaches applied to multiple myeloma

Résumé

Au cours des 2 dernières décennies, l'explosion des capacités de production de données biologiques et des connaissances sur les interactions biologiques ont permis le développement de nombreuses approches intégrant des données avec des connaissances plus générales. Notre principal objectif était de proposer de nouvelles méthodes de caractérisation et de comparaison de profils d'expressions de gènes issus de cellules cancéreuses d'individus atteints de myélome multiple et de plasmocytes normaux. Pour cela, nous proposons 2 méthodes basées sur les colorations de graphes de régulation. La première, qui permet d'inférer l'état des protéines et des facteurs de transcription à partir d'un profil d'expression, nous a permis d'identifier des activités de facteurs de transcription impliqués dans ces tumeurs. La seconde méthode permet de diviser un réseau de régulation en plusieurs sous-graphes indépendant (des *composants*) vis à vis de colorations dites "parfaites". Via cette approche, nous avons pu évaluer la similarité entre les profils d'expression et les états "parfait" des *composants* et en identifier spécifiquement perturbés dans les cellules cancéreuses associés à des voies oncogéniques.

Mots clés

Biologie des systèmes, myélome multiple, coloration de graphes, programmation par contraintes.

Abstract

During the last two decades, the huge increase of biological data production capacity and biological interactions knowledge leads to the development of many approaches integrating data and global knowledge. Our main aim was to propose new methods to characterize and compare genes expression profiles from cancer cells and normal plasma cells. For this purpose, we proposed 2 methods based on graph coloring. The first, which is able to infer the state of proteins and transcription factors from a genes expression profile, allowed us to identify transcription factor activity involved in tumors. The second method is able to identify independent subgraph (the components) based on perfect colorations. With this approach, we evaluated the similarity between genes expression profiles and "perfect states" of the components and were able to identify subgraphs specifically perturbed in cancer cells associated with oncogenic process.

Key Words

Systems biology, Multiple myeloma, graph coloring, constraint programming .