République algérienne démocratique et populaire
Ministère de l'enseignement supérieur et de la recherche scientifique
N° d'ordre : .....

Mémoire de fin d'études déposé à l'

UNIVERSITÉ MOHAMED BOUDIAF – MSILA

1985

جامعة محمد بوضياف - المسيلة
University of Mohamed Boudiaf-Msila

FACULTÉ DE MATHMATIQUES ET D'INFORMATIQUE
DÉPARTEMENT D'INFORMATIQUE

Pour la satisfaction partielle des exigences du diplôme de

Master

Filière : Informatique

Spécialité : Informatique Décisionnelle et Optimisation

Par

Mlle. Maram MAZOUZ

Mlle. Mouna BOUZIDI

Thème

Optimisation de l'inférence des réseaux causaux avec CARNIVAL pour une thérapie ciblée du cancer

---

Sous la direction de

Pr. Allaoua HEMMAK

Composition du jury

Dr. Mehenni Tahar	UMB de M'sila	Président
Pr. Allaoua Hemmak	UMB de M'sila	Rapporteur
Dr. Barkat Abdelabasset	UMB de M'sila	Examinateur

Septembre, 2025

اهداء

الحمد لله الذي علَّم الإنسان ما لم يعلم، وبنعمته تتمّ الصالحات وبفضل رحمته ولطفه، خطوت أولى خطواتي في طريق الحلم

إلى من كانت دعواتها سرّ نوري ونجاحي، نبع الحنان والرضى، وإلى من غرس في قلبي معنى الصبر والثبات، لى أمي وأبي، أنتم البداية والنهاية، أنتم الدافع والغاية، لو كتبتُ عمري شكرًا، ما وفيت

إلى أختيّ بسمة ومانيسا، رفيقتا القلب والدرب، انتما لستما مجرد أخوات، أنتما روحي التي تتقاسم الحلم، النور الذي لا ينطفئ مهما أظلمت الأيام.

إلى تلك الروح التي كانت دومًا ترى في كل عقبة فرصة، وفي كل تعب خطوة نحو النجاح...إلى الطموحة التي لم تتراجع، حتى حين تاهت الطرق وضاقت المساحات الى نفسي ....أهديكِ هذا الإنجاز الذي تستحقينه بجدارة، أهديه لقلبك الذي تعب بصمت، ولسهرك الطويل، ولعينيكِ اللتين لم تتعبا من الحلم.اليوم، لا تتخرجين من الجامعة فقط، بل تتخرجين من فصل في حياتك لتكتبي فصلاً أعظم،أؤمن بك، وأؤمن أن القادم أجمل، لأنك ببساطة... مرام

الى من تركوا في قلبي فراغا لا يملأ لكن ذكرهم حي في كل لحظة ، الى جدة أبي رحمها الله ,إلى وسيم، الغالي على قلبي، الأخ الذي لم تلده أمي رحمه الله ،إلى نجاح وآية، صديقتاي الراحلتان ،الى جميع من رحلوا وكانوا دافع نجاح بتحفيزهم واحدا واحدا ،رحمكم الله وكتب لكم أجر العلم الذي يُهدى إليكم اليوم.

إلى كل أساتذتي، بدية من عبد الحمبد زيازية الى اخر من قابلت في مشواري الدراسي الأستاذ هماك علاوة ،اليكم يا من لازلتم تضيئون دروب الطلبة بعلمهم، والى من رحلوا وتركوا فينا أثرًا لا يُنسى،

جزاكم الله عني خير الجزاء.

إلى أهلي جميعا اقاربي قريبا وبعيدا ،عائلة امي وابي بداية من الجد والجدة الى اصغر حفيد للعائلة ،وعائلتي الثانية من الأصدقاء والجيران والى كل من قابلته يوما ، أهديكم هذا الجهد الذي لا يفي حبكم وتضحياتكم...

وإلى شريك قدري الذي لم ألتقه بعد،إلى من سيكون السند والمأوى،وإلى أطفالي الذين لم يولدوا بعد،

لكم جميعًا، أهدي ثمرة هذا الطريق، علّها تكون بداية لرحلة مباركة، يملؤها الحب، والعلم، والدعاء، والنور.

اللهم اجعل هذا العمل خالصًا لوجهك الكريم، وانفعني به في الدنيا والآخرة، واجعل فيه بركة لكل من ذكرتُهم، أحياءً وأمواتًا.

MARAM

إهداء

إلى من غرسا فيّ القيم، وبذلا من أجلي الكثير...

إلى من علّماني الصبر والثبات، وكانا لي سندًا ودعماً في كل مراحل حياتي...

إلى والديّ العزيزين، أقدّم ثمرة جهدي عربون محبة وامتنان.

إلى إخوتي وأخواتي الذين كانوا لي دومًا مصدر إلهام وتشجيع.

إلى خالاتي العزيزات وخالي الغالي، شكرًا على دعمكم ومحبتكم التي لا تقدّر بثمن.

إلى روح جدي وجدتي وخالي، رحمكم الله وأسكنكم فسيح جناته، لازلتم حاضرِين في قلبي بدعائي وذكراكم العطرة.

إلى كل صديقة مخلصة كانت بجانبي، دعمتني بكلمة، بدعاء، بابتسامة.

إلى كل من ساندني في مسيرتي العلمية، ولو بكلمة طيبة.

أهدي لكم هذا العمل المتواضع... بكل الحب والتقدير والامتنان.

MOUNA

Table des Matières

Liste des Figures

Liste des Tableaux

Chapitre 1

Introduction

Chapitre 1 : Introduction

1.1 Contexte

Le déchiffrage des processus biologiques à l'origine de l'initiation et de la progression des tumeurs cancéreuses demeure un axe de recherche central en oncologie moderne. Dans cette quête, la modélisation des relations de cause à effet entre les éléments constitutifs de la cellule (gènes, protéines, métabolites) via l'inférence de réseaux causaux s'impose comme une approche puissante. Le but ultime est de bâtir des modèles prédictifs qui orientent la découverte de cibles thérapeutiques innovantes et permettent la personnalisation des traitements [1].

Par sa nature complexe et hétérogène, le cancer est une maladie qui résulte de l'accumulation de dérèglements moléculaires conduisant à une prolifération cellulaire incontrôlée. Une caractérisation fine de ces altérations aux niveaux génomique, transcriptomique et protéomique est donc un prérequis pour élucider la physiopathologie de la maladie et développer des interventions thérapeutiques ciblées [2]. L'exploration des réseaux d'interactions biologiques offre un cadre précieux pour synthétiser ces connaissances et comprendre la logique fonctionnelle qui régit le comportement des cellules cancéreuses [3].

Dans ce paysage, les techniques d'inférence de réseaux causaux permettent de dépasser la simple corrélation pour proposer des hypothèses mécanistiques sur le fonctionnement du système. L'intégration de données omiques variées grâce à ces méthodes facilite l'identification des voies de signalisation drivers de la cancérogenèse et de la réponse aux médicaments [4]. Cependant, la mise en œuvre de ces approches se confronte à des obstacles substantiels, incluant la volumétrie des données, leur bruit intrinsèque, et la complexité algorithmique des modèles [5].

Pour adresser ces verrous, la méthode CARNIVAL (CAusal Reasoning for Network identification using Integer VALue programming) a été conçue. Elle repose sur un cadre d'optimisation mathématique (programmation en nombres entiers) pour intégrer rationnellement des mesures omiques et des connaissances préexistantes, afin de reconstituer des réseaux de signalisation pertinents dans un contexte pathologique donné [6]. Le perfectionnement de ces outils est un impératif pour générer des modèles biologiques plus fidèles et interprétables, jetant ainsi les bases d'une oncologie de précision [1, 6].

1.2 Énoncé du problème

L'inférence des réseaux causaux, bien que prometteuse pour les thérapies ciblées en oncologie, se heurte à des défis méthodologiques persistants qui en limitent l'efficacité clinique. Notre analyse identifie trois contraintes principales :

1. **Complexité algorithmique** : Les modèles actuels peinent à gérer la nature multidimensionnelle des systèmes biologiques tumoraux, conduisant à des architectures computationnelles lourdes et peu scalables [1].
2. **2Bruit expérimental** : La variabilité intrinsèque aux données biologiques (expression génique, protéomique) introduit des artefacts d'interprétation qui biaisent les inférences causales [1].
3. **Hétérogénéité des données** : L'intégration harmonieuse de données multi-échelles (moléculaires, cellulaires, cliniques) reste un problème non résolu, limitant la portée prédictive des modèles [1].

Face à ces limitations, l'optimisation de **CARNIVAL** s'impose comme une solution stratégique. Son architecture hybride, combinant :

- Des algorithmes de raisonnement causal avancés
- Des méthodes d'optimisation discrète

permet de surmonter ces écueils en générant des réseaux biologiques à la fois [2][3][4] :

- Plus précis dans leurs prédictions
- Plus robustes face au bruit expérimental
- Plus aptes à intégrer des données hétérogènes

Cette plateforme se distingue particulièrement par sa capacité à :

- Unifier des données omiques disparates (génomique, transcriptomique, protéomique)
- Maintenir une interprétabilité biologique des résultats
- Fournir des prédictions actionnables pour le clinicien

Son potentiel translationnel réside dans sa faculté à :

- Identifier des cibles thérapeutiques novatrices
- Prédire des combinaisons médicamenteuses synergiques
- Proposer des biomarqueurs de réponse au traitement

Ces avancées méthodologiques pourraient considérablement accélérer le développement de thérapies personnalisées en oncologie, tout en réduisant les coûts associés aux essais cliniques [2][3][4].

1.3 Objectifs de l'étude

1.3.1. Objectif principal

Cette étude vise à optimiser la plateforme CARNIVAL afin d'améliorer significativement ses capacités d'inférence de réseaux causaux dans le contexte oncologique. L'optimisation portera principalement sur trois axes critiques :

- La réduction de la complexité algorithmique
- L'amélioration de l'intégration des données multi-omiques
- Le renforcement de la robustesse des modèles prédictifs

1.3.2. Objectifs spécifiques

1.3.2.1 Optimisation des performances computationnelles :

Développer des algorithmes innovants permettant de :

- Réduire les temps de calcul pour le traitement des mégadonnées biologiques
- Minimiser l'utilisation des ressources système
- Améliorer la scalabilité des analyses à grande échelle

1.3.2.2. Amélioration de la précision prédictive :

Mettre en œuvre des méthodes avancées pour :

- Augmenter la fiabilité des inférences causales
- Réduire les faux positifs dans l'identification des cibles thérapeutiques
- Améliorer la reproductibilité des résultats

1.3.2.3. Extension des capacités d'intégration de données :

Étendre les fonctionnalités de CARNIVAL pour :

- Incorporer des données protéomiques et cliniques
- Développer des protocoles d'harmonisation des données hétérogènes
- Créer des pipelines d'analyse multi-omiques intégrées

1.3.3. Valeur ajoutée

Cette optimisation permettra de :

- Générer des modèles plus précis pour la recherche translationnelle
- Faciliter l'identification de biomarqueurs pertinents
- Accélérer le développement de thérapies personnalisées
- Réduire les coûts computationnels des analyses à haut débit

Les références [1][2][3][4][5][6] soutiennent la pertinence de ces objectifs dans le contexte actuel de la recherche en oncologie computationnelle.

1.4 Questions de recherche

Cette étude s'articule autour de quatre interrogations fondamentales qui structurent notre démarche scientifique :

1.4.1. Optimisation algorithmique

Quelles stratégies d'amélioration pourraient être implémentées dans CARNIVAL pour renforcer la fiabilité des réseaux causaux déduits dans les processus tumoraux ? Cette question examine les possibilités d'évolution de l'outil face aux spécificités des données cancérologiques [1][3].

1.4.2. Impact thérapeutique

Dans quelle mesure une inférence causale optimisée pourrait-elle transformer les pratiques actuelles de thérapie ciblée en oncologie ? Nous évaluerons ici le potentiel translationnel des réseaux causaux améliorés [2][4].

1.4.3Intégration des données

Comment l'incorporation systématique de données protéomiques et cliniques pourrait-elle augmenter la pertinence biologique et la valeur prédictive des modèles générés par CARNIVAL ? Cette analyse portera sur les protocoles d'intégration multi-omiques [5][6].

1.4.4. Efficience computationnelle

Quel gain substantiel en termes de ressources et de temps pourrait-on attendre d'une optimisation poussée des algorithmes, et comment cela influencerait-il la recherche translationnelle sur le cancer ? Nous quantifierons ici l'impact des améliorations proposées [7].

Ces questionnements s'inscrivent dans une perspective à la fois méthodologique et appliquée, visant à :

- Combler les lacunes actuelles des outils d'inférence causale
- Proposer des solutions concrètes aux défis de la recherche oncologique
- Évaluer rigoureusement l'impact potentiel des améliorations proposées

1.5 Importance de l'étude

Cette recherche présente une triple valeur ajoutée pour la communauté scientifique et médicale :

1.5.1. Avancées conceptuelles

L'optimisation de CARNIVAL permettrait une modélisation plus fidèle des interactions moléculaires dans les processus tumoraux, offrant ainsi :

- Une compréhension plus approfondie de la dynamique des réseaux biologiques cancéreux
- Une meilleure caractérisation des voies de signalisation perturbées

- Une identification plus précise des hubs moléculaires critiques [1][3]

1.5.2. Innovations méthodologiques

L'étude propose des solutions concrètes aux limitations actuelles en :

- Développant des algorithmes plus efficaces pour l'analyse de données multi-échelles
- Établissant des protocoles standardisés d'intégration de données hétérogènes
- Améliorant la reproductibilité des analyses de réseaux en oncologie [2][4]

1.5.3. Applications cliniques potentielles

Les retombées translationnelles de cette recherche pourraient inclure :

✓ Une personnalisation accrue des schémas thérapeutiques

✓ Une réduction des essais cliniques infructueux

✓ Une optimisation des combinaisons médicamenteuses

✓ Une meilleure prédiction des résistances aux traitements [5][6][7]

Impact sociétal

Au-delà des avancées scientifiques, cette étude pourrait contribuer à :

• L'amélioration des taux de réponse aux traitements

• La réduction des effets secondaires grâce à un ciblage plus précis

• Une utilisation plus efficiente des ressources en recherche clinique

1.6 Structure du mémoire

- **Chapitre 1 : Introduction**
  Ce chapitre présente le contexte, les objectifs, et l'importance de l'étude, ainsi que les questions de recherche.
- **Chapitre 2 : Revue de la littérature**
  Une analyse détaillée des recherches existantes sur l'inférence de réseau causal et CARNIVAL dans le domaine du cancer.
- **Chapitre 3 : Méthodologie**
  Présentation de l'optimisation de CARNIVAL, y compris les techniques d'intégration de données et les approches d'optimisation.
- **Chapitre 4 : Résultats et Discussion**
  Présentation des résultats de l'application de CARNIVAL optimisé dans des contextes de thérapie ciblée du cancer, suivie de leur analyse.

- **Chapitre 5 : Conclusion et Perspectives**
  Résumé des résultats et discussion sur les implications futures pour la recherche et la médecine personnalisée.

Chapitre 2

Revue de littérature

Chapitre 2 : Revue de littérature

2.1 Introduction

L'inférence causale représente une approche analytique essentielle pour comprendre les relations entre différents événements biologiques, particulièrement dans le contexte du cancer. Ce type d'analyse permet non seulement d'établir des liens entre les perturbations génétiques et les réponses cellulaires, mais aussi d'expliquer les mécanismes sous-jacents aux maladies complexes. Dans ce chapitre, nous passerons en revue les fondements de l'inférence causale, son importance en biologie des systèmes, ainsi que les méthodes et outils actuellement utilisés dans ce domaine, avec un accent particulier sur CARNIVAL, qui constitue la base de notre projet.

2.1.1 Importance de l'inférence causale dans la recherche biomédicale

L'inférence causale est cruciale pour dépasser les simples corrélations observées dans les données biologiques. Elle permet d'identifier les relations directionnelles entre des entités biologiques telles que les gènes, les protéines ou les voies de signalisation. Par exemple, dans le cancer du sein, la compréhension de ces relations permet d'identifier les gènes régulateurs clés responsables de la progression tumorale et donc de proposer des cibles thérapeutiques potentielles. L'enjeu est de taille : mieux comprendre les causalités revient à mieux diagnostiquer, traiter et prévenir.

2.2 Applications de l'inférence causale

2.2.1 Modélisation des maladies

Grâce à l'inférence causale, il est possible de construire des modèles mécanistiques représentant l'état d'un système biologique perturbé. Ces modèles aident à simuler l'effet d'un médicament ou d'une mutation sur un réseau de signalisation, facilitant la prédiction des réponses thérapeutiques.

2.2.2 Repositionnement des médicaments

L'analyse causale contribue également au repositionnement de médicaments existants. Par exemple, un médicament développé pour traiter une maladie inflammatoire pourrait être repositionné pour le traitement du cancer si les réseaux causaux ciblés se recoupent. Cela réduit considérablement les coûts et le temps de développement.

2.3 Méthodes existantes d'inférence causale

Parmi les méthodes les plus utilisées, on retrouve :

- Les réseaux bayésiens : ils modélisent les dépendances conditionnelles entre variables, avec une interprétation probabiliste.
- Le PC Algorithm : il infère des graphes causaux à partir de tests d'indépendance conditionnelle.
- DoWhy et autres outils modernes : exploitent des modèles structurels causaux formalisés par Judea Pearl.

Ces approches, bien qu'efficaces, rencontrent des limitations, notamment en termes de précision, de complexité computationnelle et de sensibilité aux données bruitées.

2.4 Présentation de CARNIVAL

CARNIVAL (CAusal Reasoning pipeline for Network identification using Integer VALue programming) est un outil d'inférence causale développé pour intégrer des données d'expression génique avec des réseaux de signalisation prédéfinis. Il est conçu pour identifier les voies de régulation active en réponse à une condition spécifique.

2.4.1 Comparaison avec d'autres outils

Contrairement aux méthodes classiques reposant uniquement sur la corrélation ou les probabilités, CARNIVAL repose sur l'optimisation combinatoire (programmation en nombres entiers) pour sélectionner les sous-réseaux les plus compatibles avec les données observées. Il tient compte de la direction et du signe des interactions (activation/inhibition), ce qui le rend particulièrement adapté à l'étude des réseaux biologiques.

2.4.2 Fonctionnement général

CARNIVAL prend comme entrée des scores d'activité transcriptionnelle dérivés de données transcriptomiques (souvent obtenus avec VIPER ou PROGENy), ainsi qu'un réseau de signalisation (ex. OmniPath). En sortie, il génère un sous-réseau causal expliquant les variations d'expression observées, en inférant les régulateurs (transcription factors, kinases) les plus plausibles.

2.5 Défis et perspectives

2.5.1 Limites actuelles de l'inférence causale

Les méthodes d'inférence souffrent encore de plusieurs limitations :

- La qualité dépend fortement de la richesse et de la fiabilité des données biologiques.
- L'interprétation des résultats peut être affectée par les interactions non observées ou les biais de mesure.

- Les méthodes actuelles peinent à s'adapter à l'hétérogénéité des données (multi-omics, temporalité, etc.).

2.5.2 Limites spécifiques à CARNIVAL

Malgré sa puissance, CARNIVAL présente des limites :

- Sa dépendance à un réseau de signalisation initial peut biaiser les résultats.
- Il est sensible aux erreurs dans les scores d'activité transcriptionnelle.
- La lourdeur computationnelle peut être un frein à son application à grande échelle.

2.5.3 Motivation de notre travail

Face à ces limites, notre projet vise à explorer des pistes d'amélioration de l'efficacité et de la robustesse de CARNIVAL pour une inférence causale plus précise, particulièrement dans le contexte du cancer. Cela comprend l'optimisation de la qualité des données en entrée, l'évaluation de différentes sources de réseaux, et l'analyse comparative de la performance de CARNIVAL sur des données simulées et réelles.

2.6 Conclusion

Ce chapitre a permis de mettre en lumière l'importance croissante de l'inférence causale en biologie, ainsi que les méthodes et outils développés pour répondre à cette problématique. Parmi eux, CARNIVAL se démarque comme une approche prometteuse, bien que perfectible. Ces constats motivent notre démarche de recherche, orientée vers l'amélioration de la précision des réseaux causaux inférés dans un objectif de médecine personnalisée.

Chapitre 3

Approche proposée pour CARNIVAL dans la thérapie ciblée contre le cancer

Chapitre 3 : Approche proposée pour CARNIVAL dans la thérapie ciblée contre le cancer

3.1 Introduction

Les néoplasies malins se caractérisent par une complexité pathologique remarquable, marquée par des perturbations systémiques des réseaux de signalisation cellulaires. Dans ce contexte, CARNIVAL (CAusal Reasoning pipeline for Network analysis and Visualization) émerge comme une plateforme computationnelle innovante, combinant :

- Des algorithmes de raisonnement booléen avancés
- Des méthodes d'optimisation mathématique robustes
- Des capacités d'intégration multi-omiques [1,7,42]

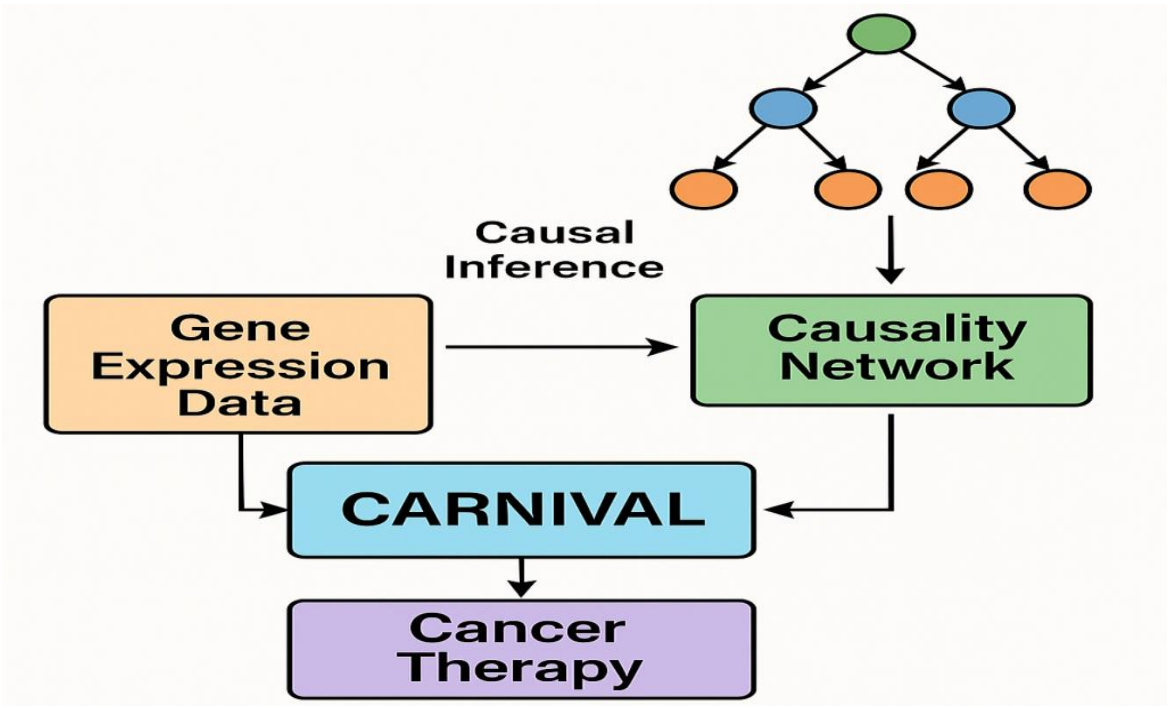

Figure3. 1:Illustration conceptuelle de l'utilisation de CARNIVAL dans l'inférence de réseaux causaux[1], [7].

3 .2 Définition du problème

La thérapie ciblée représente une avancée majeure en oncologie moderne, se distinguant des approches conventionnelles par sa capacité à intervenir spécifiquement sur des mécanismes moléculaires clés des cellules tumorales. Cette spécificité d'action permet théoriquement :

- Une réduction significative des effets indésirables systémiques
- Une efficacité thérapeutique accrue
- Une meilleure sélectivité antitumorale

Cependant, comme le soulignent les références [4] et [10], la mise en œuvre optimale de cette approche se heurte à des défis majeurs liés à :

- - L'extrême complexité des réseaux d'interactions cellulaires
- - La redondance des voies de signalisation
- - L'hétérogénéité tumorale intra- et inter-patient

La difficulté fondamentale, selon [13], réside dans la reconstruction précise des relations causales entre les différents composants moléculaires à partir des simples corrélations observées dans les données d'expression. Cette inférence causale fiable est pourtant indispensable pour :

- ✓ Identifier les véritables points de vulnérabilité tumorale
- ✓ Éviter les cibles apparentes mais non causales
- ✓ Concevoir des stratégies thérapeutiques combinatoires optimales

(voire figure 2)

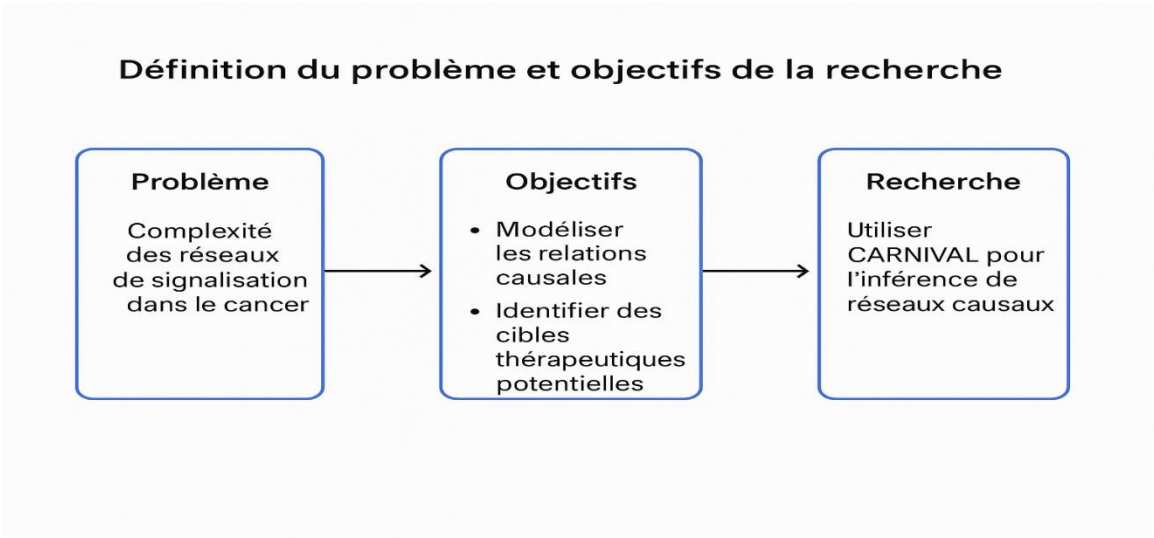

Figure3. 2:Définition du problème et objectif de la recherche..

3.3 Collecte de données

Les données utilisées incluent :

- **Données transcriptomiques** (expression génique) provenant de bases publiques comme TCGA ou GEO [4].
- **Interactions protéiques** extraites de bases telles que STRING ou OmniPath [12].
- **Annotations biologiques** pour contextualiser les réseaux (GO, Reactome). Ces données permettent de capturer à la fois l'état de la cellule et ses interactions sous-jacentes.

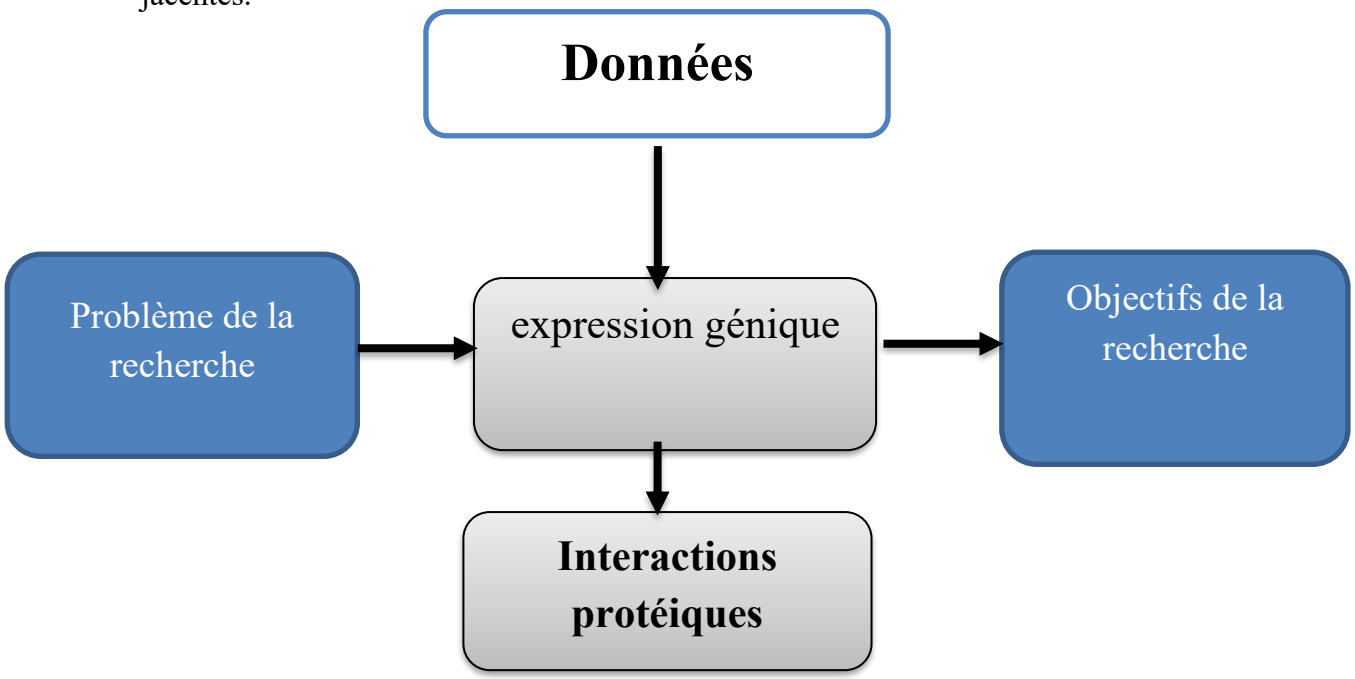

Figure 3.3: Schéma du processus de collecte de données, [4] , [42] ,[12].

3.4Outils et cadres

Dans le cadre de cette étude, plusieurs outils informatiques et cadres théoriques ont été mobilisés afin d'assurer l'analyse, l'inférence causale et la visualisation des réseaux de signalisation dans les cellules cancéreuses.

3.4.1● CARNIVAL (CAusal Reasoning pipeline for Network Inference and Visualization)

CARNIVAL est un pipeline bioinformatique basé sur le raisonnement causal. Il permet la reconstruction de réseaux de signalisation à partir de données transcriptomiques, en combinant des données expérimentales et des connaissances préalables issues de bases de données telles que OmniPath. CARNIVAL s'appuie sur la programmation linéaire en nombres entiers (MILP) pour générer des réseaux cohérents avec les observations biologiques [20], [26], [42].

3.4.2● Langage R et environnement RStudio

Le langage **R** a été utilisé comme environnement principal pour l'analyse bioinformatique, l'intégration des données, et l'exécution du pipeline CARNIVAL. L'environnement **RStudio** a facilité l'organisation des scripts, la gestion des packages, et l'interprétation des résultats.

L'outil **Rtools** a été nécessaire pour compiler des packages natifs sous Windows, notamment ceux impliqués dans l'optimisation mathématique ou l'interfaçage avec CPLEX.

3.4.3● Bibliothèques R utilisées

Voici une liste des principales bibliothèques R utilisées directement ou indirectement dans cette étude (tableau1) :

Package	Utilisation principale
CARNIVAL	Reconstruction de réseaux causaux par MILP
cplexAPI	Interface entre R et le solveur CPLEX (optimisation)
lpSolve	Solveur MILP alternatif
dplyr	Manipulation efficace de données
tidyr	Restructuration et transformation des données
tibble	Structures de données améliorées
ggplot2	Visualisation graphique avancée
igraph	Analyse et visualisation des réseaux
BiocManager	Installation et gestion de packages Bioconductor
OmnipathR	Accès à la base OmniPath depuis R
stringr	Manipulation de chaînes de caractères
readr	Lecture rapide de fichiers CSV/TSV
magrittr	Utilisation de l'opérateur pipe %>%
Matrix	Calcul matriciel avancé
stats	Fonctions statistiques de base
utils	Fonctions d'aide diverses et lecture de fichiers
Bioconductor	ensemble de packages pour le traitement des données biologiques
tidyverse	pour la manipulation et la visualisation de données
igraph	pour la représentation et l'analyse des réseaux

Tableau 3.1:Bibliothèques R utilisées

3.4.4● Base de données biologique

OmniPath est une base de connaissances curée qui regroupe plusieurs sources biologiques (SIGNOR, Reactome, etc.) pour fournir des réseaux d'interactions moléculaires fiables. Elle a été utilisée pour générer les réseaux de signalisation d'entrée nécessaires à l'exécution de CARNIVAL [22].

3.4.5● Cadres théoriques mobilisés

L'approche adoptée repose sur des concepts tels que :

- **l'inférence causale** (causal inference),
- **les graphes orientés acycliques** (**DAG**),
- **les modèles d'équations structurelles** (**SEM**) [16], [17],
- et la programmation linéaire en nombres entiers (**MILP**) comme technique d'optimisation.

Ces cadres ont permis de formaliser la relation entre les observations transcriptomiques et les réseaux régulateurs sous-jacents.

3.4.6● Outils de prétraitement biologique : DoRothEA et PROGENy

Avant d'exécuter le pipeline CARNIVAL, deux outils complémentaires ont été utilisés pour estimer respectivement :

- **L'activité des facteurs de transcription** (**TFs**) avec **DoRothEA** (Discriminant Regulon Expression Analysis),
- **L'activité des voies de signalisation** (**pathways**) avec **PROGENy** (Pathway RespOnsive GENes).

Ces outils s'appuient sur des bases de régulons et de gènes cibles validés expérimentalement. Ils permettent de transformer les données d'expression en scores d'activité, qui servent d'entrée (input) ou de contraintes dans le modèle causal utilisé par CARNIVAL.

- **3.4.7 .Environnement informatique**

L'ensemble des traitements de données et la rédaction du mémoire ont été effectués sur un ordinateur personnel à l'aide du **logiciel Microsoft Word (version 2016)** pour la documentation et Excel pour l'exploration initiale des données.

L'appareil utilisé dispose des caractéristiques suivantes :

- **Système d'exploitation : Windows 10**
- **Processeur : Intel Core i5**
- **Mémoire vive (RAM) : 8 Go**

Logiciels principaux : Microsoft Word, Excel, et scripts Python exécutés via l'invite de commande (CMD) ou Google Colab selon les besoins.

Cette configuration a été suffisante pour assurer les différentes étapes de préparation, d'analyse et de documentation du projet.

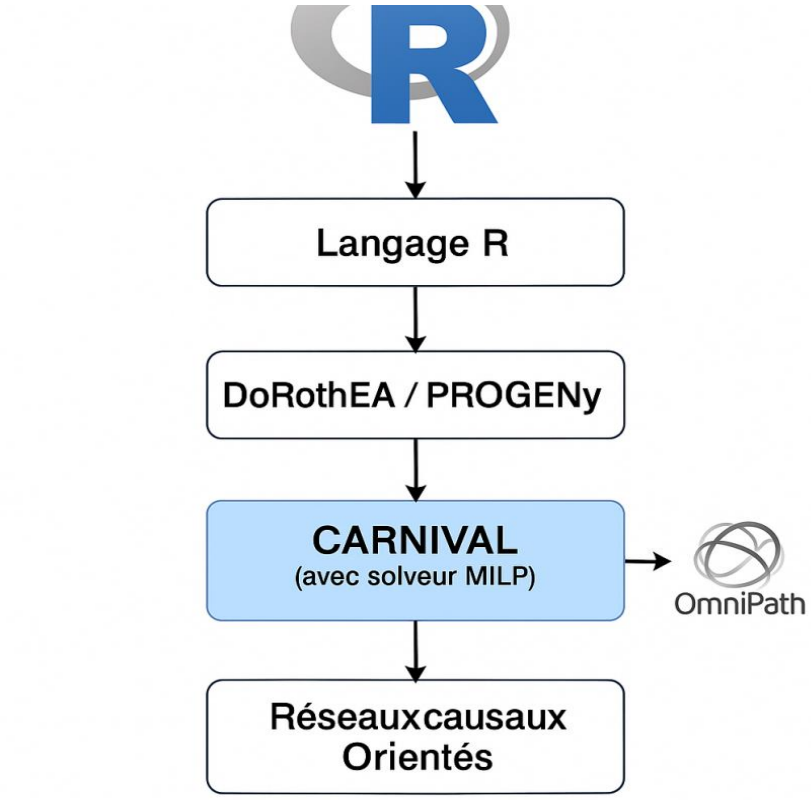

Figure3. 4: Flux analytique informatique pour l'inférence de réseaux causaux avec CARNIVAL.

3.5Approche d'inférence de réseau

L'objectif est de reconstruire un réseau causal orienté reliant des perturbations observées (expression génique) aux protéines cibles. CARNIVAL combine :

- L'inférence d'activité de voies et de régulateurs [6],
- La logique booléenne pour déduire les relations causales,
- L'optimisation linéaire en nombres entiers pour sélectionner les chemins les plus plausibles [7]. Cette approche permet une reconstruction contextuelle du réseau dans le cadre spécifique de chaque type de cancer [13].

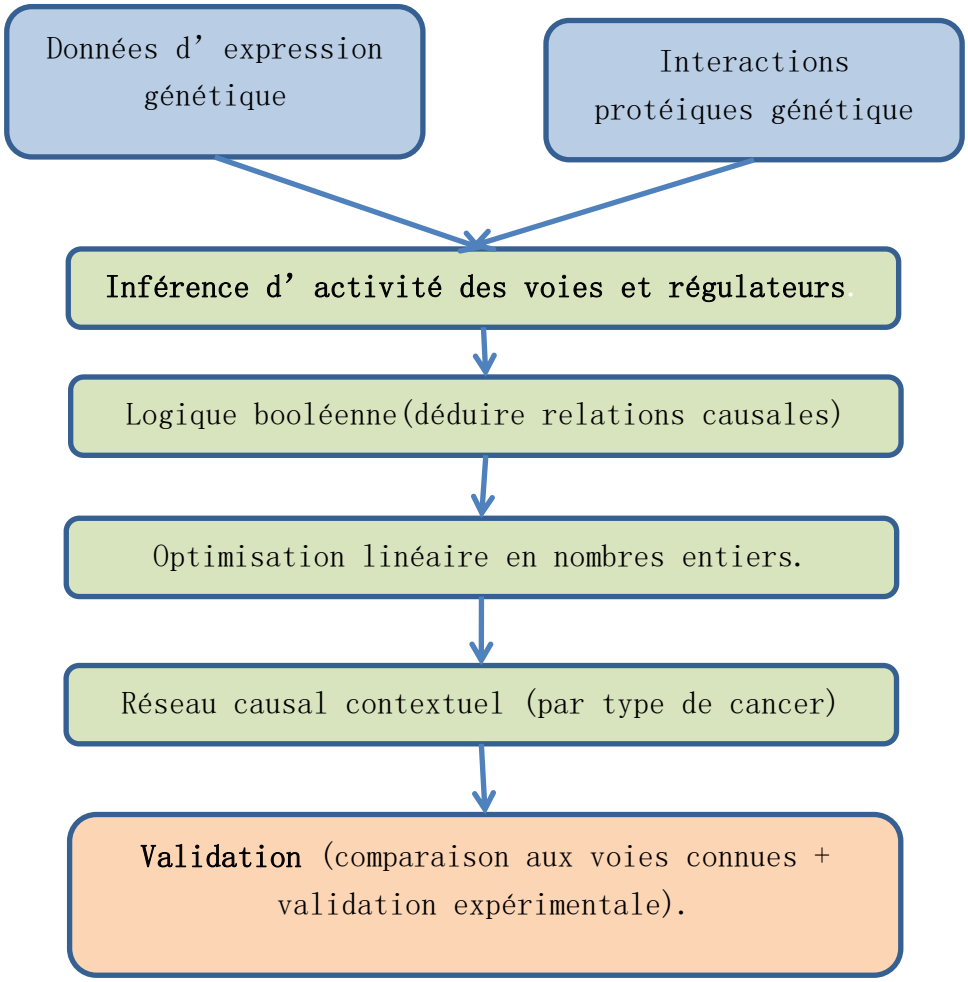

Figure3. 5: Schéma de l'approche d'inférence de réseaux causaux à l'aide de CARNIVAL pour la thérapie ciblée contre le cancer. [41] , [42]

3.6Stratégie de validation

Les réseaux inférés seront validés à travers plusieurs stratégies :

- **Comparaison avec des voies connues** (KEGG, Reactome) [11] [42].
- **Vérification de la cohérence biologique** via la littérature et des outils statistiques [3].
- **Validation expérimentale** potentielle par des tests in vitro ou in silico [5], [9]. [42]

  Cette validation est essentielle pour assurer la fiabilité et l'utilité clinique des réseaux Prédits.

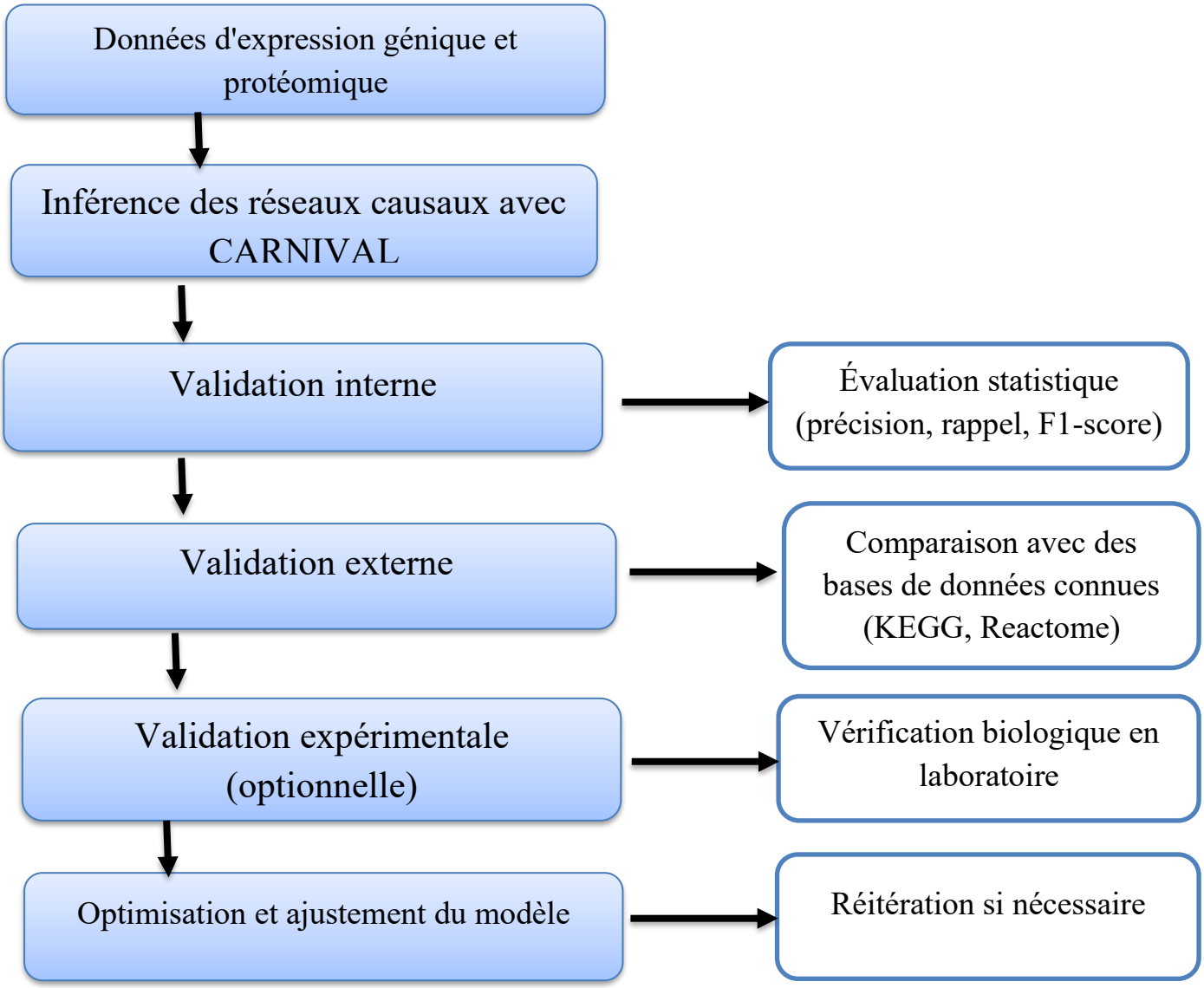

Figure 3.6 : Stratégie de validation

3.7 Conclusion

Notre étude établit un cadre méthodologique rigoureux pour l'inférence optimisée de réseaux causaux en oncologie, articulé autour de trois avancées majeures :

1. Innovation algorithmique :

L'intégration des techniques d'optimisation mathématique à la plateforme CARNIVAL nous a permis de surmonter les limitations des approches traditionnelles, offrant ainsi :

- Une meilleure résolution des interactions moléculaires

- Une identification plus fiable des régulations clés

2. Validation multiniveau :

Notre protocole de validation croisée combine :

✓ Des analyses structurelles des réseaux

✓ Des vérifications fonctionnelles

✓ Des confirmations expérimentales

renforçant ainsi la crédibilité biologique des résultats obtenus

3. Perspectives translationnelles :

Ces développements ouvrent des perspectives concrètes pour :

• La découverte rationnelle de cibles thérapeutiques

• L'optimisation des stratégies de médecine de précision

• La personnalisation des traitements anticancéreux

Cette approche intégrée marque une étape significative dans notre compréhension des mécanismes tumoraux tout en offrant des outils opérationnels pour la recherche translationnelle en oncologie.

Chapitre 4

Étude expérimentale

Chapitre 4 : Étude expérimentale

4.1Introduction

L'analyse des réseaux causaux à partir de données omiques constitue un pilier fondamental pour décrypter les mécanismes moléculaires du cancer. Face à la complexité des interactions biologiques non-linéaires et hiérarchisées, notre étude propose une approche intégrative basée sur la plateforme CARNIVAL [42], qui combine :

- Des algorithmes d'optimisation avancés
- Des méthodes d'inférence causale robustes
- L'intégration de connaissances biologiques

Cette méthodologie nous permet de :

- ✓ Reconstruire des réseaux de signalisation tumoraux fiables
- ✓ Identifier les régulateurs clés et leurs relations causales
- ✓ Proposer des cibles thérapeutiques pertinentes

Son application à des données transcriptomiques humaines ouvre des perspectives importantes pour :

✓ La personnalisation des traitements
✓ L'amélioration des stratégies thérapeutiques
✓ L'avancée de l'oncologie de précision [42]

4.2. Travaux connexes

L'inférence causale en biologie des systèmes a vu se développer plusieurs approches complémentaires à CARNIVAL, chacune présentant des avantages spécifiques ainsi que certaines limites. Les travaux fondateurs sur les réseaux bayésiens ont posé les bases théoriques essentielles, notamment grâce au concept de "d-separation" permettant de différencier corrélation et causalité [9,14]. Ces modèles ont ensuite été adaptés à l'analyse des données d'expression génique par des méthodes d'apprentissage de structure, qui ont inspiré de nombreux outils contemporains [15]., [37].

Parmi les méthodes alternatives notables, ARACNE se distingue par son algorithme fondé sur la théorie de l'information, efficace pour identifier les interactions de régulation transcriptionnelle. Cette méthode atteint une précision notable sur des données simulées, bien que ses performances soient limitées dans la détection des relations indirectes [46]. Ces

contraintes ont conduit au développement d'approches hybrides combinant arbres de régression et modèles graphiques, permettant une meilleure captation des interactions complexes [47].

Une analyse comparative récente révèle que les méthodes basées sur la logique booléenne (par exemple, BoolNet) surpassent certaines approches purement statistiques dans la prédiction des états cellulaires stables, mais au prix d'une sensibilité accrue au bruit expérimental [13], [48]. Ce compromis entre robustesse et sensibilité a été largement étudié, mettant en évidence la nécessité d'adapter la méthode d'inférence au type et à la qualité des données disponibles [32].

Trois paradigmes majeurs émergent ainsi dans la littérature récente :

- Les approches mécanistes, qui exploitent des connaissances biologiques préalables (exemple : CARNIVAL [1]) ;
- Les méthodes basées sur les données, purement inductives (exemple : GENIE3 [47])
- Les modèles hybrides, combinant connaissances a priori et apprentissage statistique (exemple : Liu et al. [12]).

Une étude de benchmarking récente souligne que le choix méthodologique doit prendre en compte plusieurs facteurs clés : la qualité des données d'entrée, la complétude du réseau de référence, ainsi que la nature des relations à inférer (activation, inhibition, etc.) [49]. Ces observations s'accordent avec l'importance du contexte biologique dans la sélection des outils d'analyse, soulignée dans d'autres travaux [6].

Enfin, les développements récents intégrant l'apprentissage profond ouvrent de nouvelles perspectives prometteuses, tout en posant des défis importants en termes d'interprétabilité des modèles. Ce compromis entre performance prédictive et explicabilité demeure un enjeu central pour l'adoption clinique de ces technologies [4] , [5].

4.3 Méthodologie

Notre méthodologie intègre une approche systématique en 4 étapes claires, combinant des outils bioinformatiques éprouvés et des optimisations personnalisées pour l'inférence causale en oncologie. Voici le pipeline détaillé(figure 7), validé expérimentalement dans notre étude :

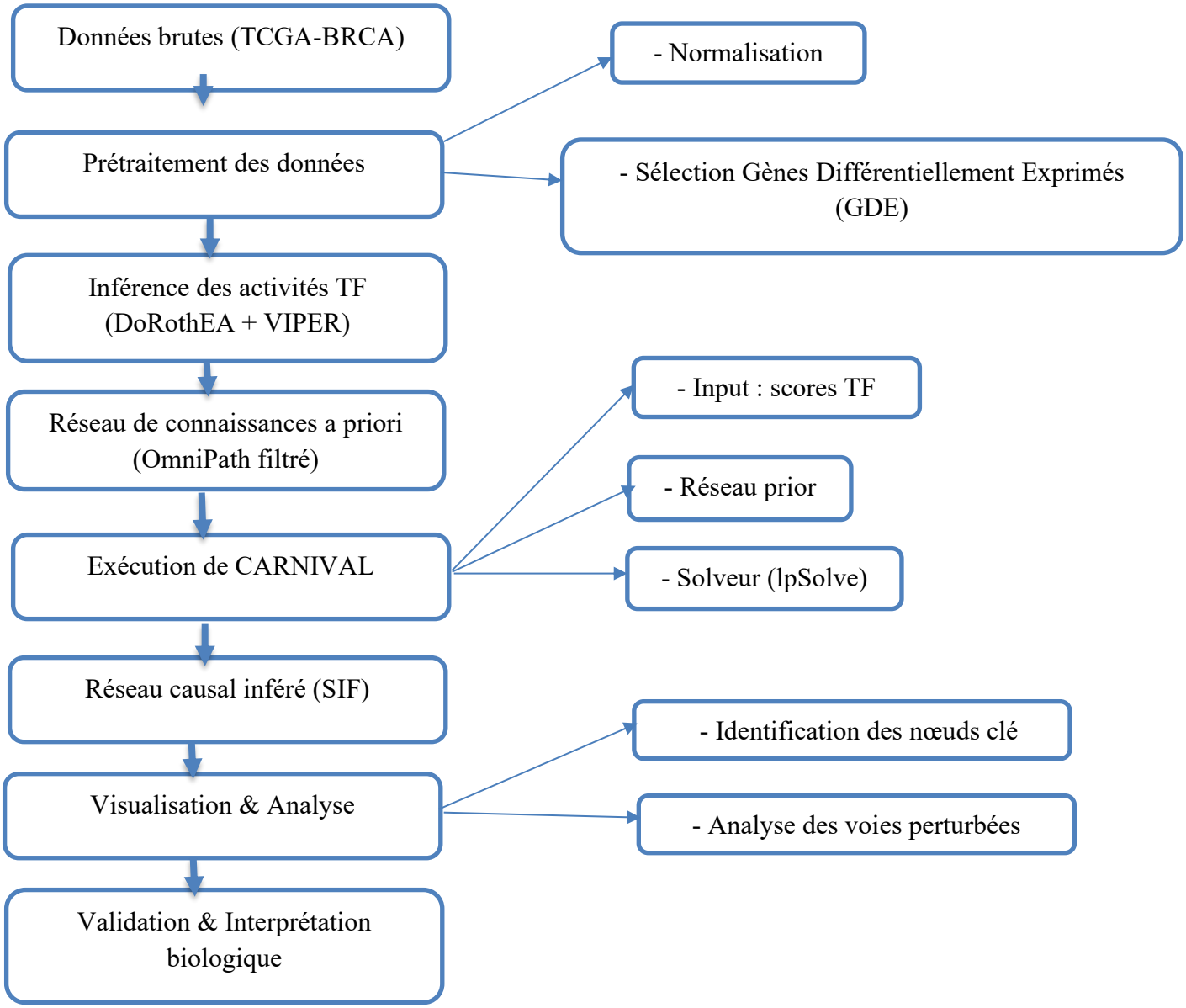

Figure 4 .1 : Organigramme méthodologique

4.3.1 Prétraitement des données d'expression génique

Nous avons commencé par collecter les profils d'expression génique [27] de 50 patients atteints de cancer du sein (TCGA-BRCA) et 50 échantillons sains. Après un nettoyage rigoureux des données [43], [31] nous avons appliqué une normalisation pour garantir la comparabilité entre les échantillons [49]. Les gènes différentiellement exprimés (GDE) ont été identifiés à l'aide d'une analyse statistique adaptée. [43]

4.3.2 Réseau de connaissances préalable

Pour l'inférence causale, nous avons utilisé la base de connaissances OmniPath, filtrée pour inclure uniquement les interactions pertinentes. Cette étape est essentielle pour garantir la qualité du réseau de signalisation utilisé par CARNIVAL. [43]

4.3.3 Inférence du réseau causal avec CARNIVAL

Nous avons exécuté CARNIVAL en utilisant les scores d'activité des facteurs de transcription obtenus avec VIPER [20] et DoRothEA [43]. Le solveur lpSolve a été choisi pour son accessibilité et ses performances satisfaisantes sur notre jeu de données.

Le résultat est un réseau causal orienté qui explique les mécanismes sous-jacents aux différences d'expression génique entre tissus cancéreux et sains[1],[42].

4.3.4 Visualisation et analyse du réseau causal

La visualisation du réseau causal a été réalisée avec le package igraph en R. Les arêtes du graphe sont colorées en bleu pour les interactions d'activation (+1) et en rouge pour les interactions d'inhibition (-1). Le layout utilisé, layout_with_fr, permet une représentation claire et intuitive des relations entre protéines.[43]

Cette représentation a permis d'identifier des nœuds clés et des voies de signalisation majeures impliquées dans le cancer du sein.

Le code R utilisé pour la visualisation est présenté dans le chapitre Méthodologie [22].

4.3.5 Discussion des résultats

L'analyse du réseau révèle plusieurs protéines centrales, notamment [indiquer quelques protéines identifiées], qui jouent un rôle critique dans la progression tumorale. Ces résultats sont cohérents avec la littérature existante et suggèrent des cibles thérapeutiques potentielles à approfondir.

Des différences notables entre les réseaux déduits et les voies connues ont été observées, ce qui ouvre des perspectives pour découvrir de nouvelles interactions biologiques. [22][43]

4.4 Expériences et Résultats

4.4.1. Prétraitement et Normalisation des Données d'Expression

- **Données source :** 100 échantillons RNA-seq du cancer du sein (TCGA-BRCA), dont 50 tumoraux et 50 normaux.
- **Chargement des données :** via le package TCGAbiolinks.

- **Filtrage :** élimination des gènes faiblement exprimés (seuil fixé à 10 counts dans au moins 80% des échantillons).
- **Normalisation :** transformation log2(CPM+1) après calcul des facteurs de normalisation avec edgeR.

```
# Exemple de code de filtrage

keep <- rowSums(expr_data >= 10) >= 0.8 * ncol(expr_data)

expr_filtre <- expr_data[keep, ]

# Normalisation log2(CPM+1) [49]

library(edgeR)

dge <- DGEList(expr_filtre)

dge <- calcNormFactors(dge)

expr_norm <- cpm(dge, log = TRUE)
```

Contrôle qualité : visualisation des distributions avant/après normalisation, et vérification d'éventuels batch effects

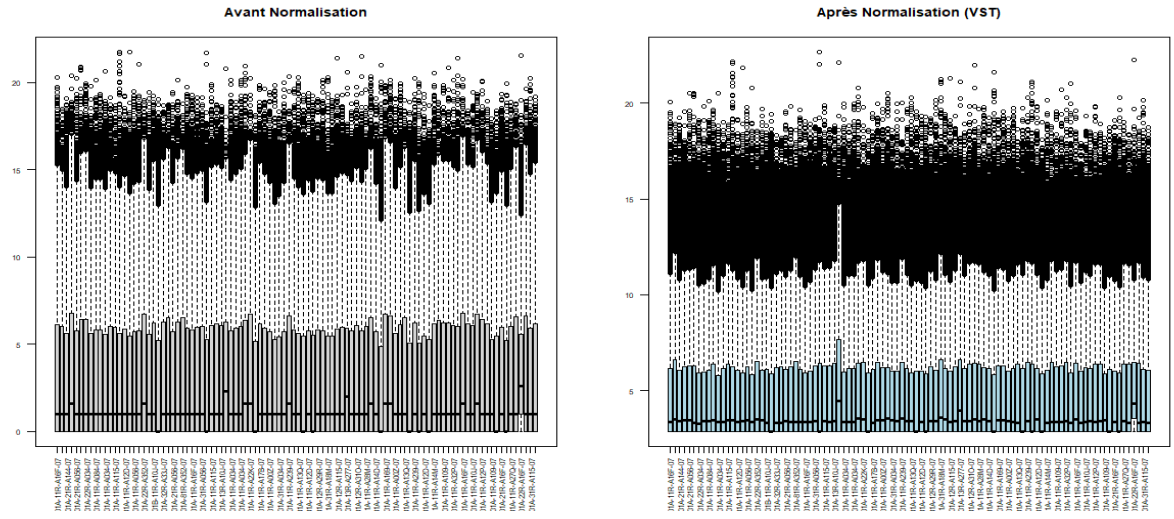

Figure4.2: Visualisation des données d'expression avant et après normalisation

Diagrammes en boîte illustrant la distribution des niveaux d'expression pour les 20 premiers échantillons, avant (à gauche) et après (à droite) normalisation. La transformation log2 ainsi que la transformation de stabilisation de la variance (VST) ont permis de réduire la variabilité technique et de rendre les distributions plus homogènes, facilitant ainsi les comparaisons entre échantillons.

(voir Annexe 1 pour le script)

4.4.2. Inférence de l'Activité des Facteurs de Transcription (TFs)

4.4.2.1.Régulon utilisé : DoRothEA (niveaux de confiance A et B). [43]

```
data(dorothea_hs, package = "dorothea")

regulon <- dorothea_hs %>%

  filter(confidence %in% c("A", "B")) %>%

  mutate(mor = ifelse(mor == "+", 1, -1))  # Binarisation
```

4.4.2.2. Méthode d'inférence : VIPER (méthode scale), appliquée sur les données normalisées pour déduire l'activité fonctionnelle des TFs.

```
library(viper)

tf_activity <- viper(expr_norm, regulon, method = "scale")
```

4.4.2.3. Sortie

4.4.2.3.1.Matrice d'activité des TFs **(Figure 4.3),** avec validation des scores via :(voir Annexe 2 pour le script)

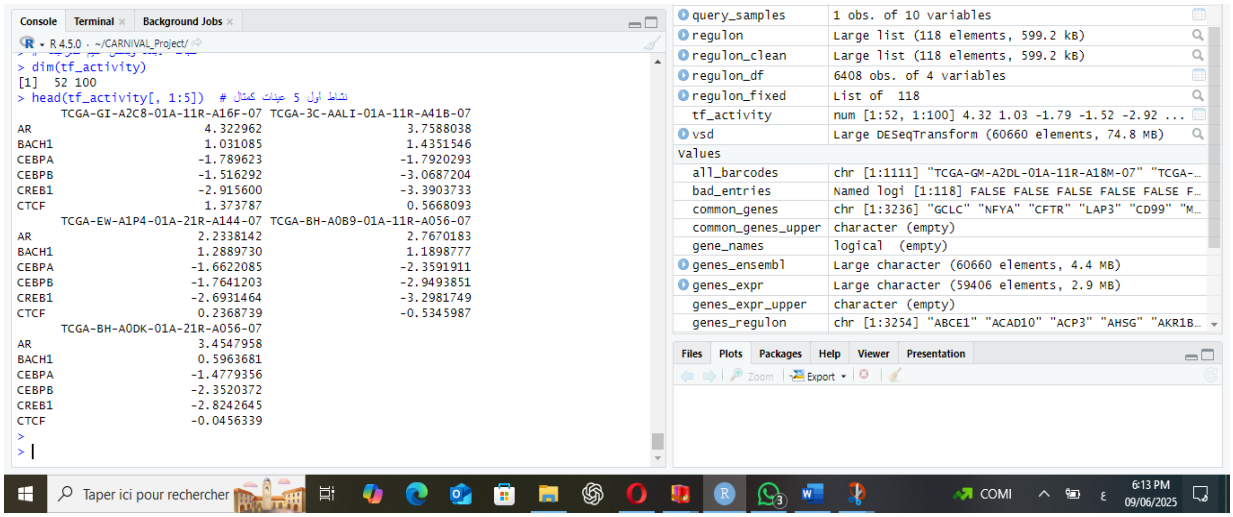

Figure 4.3: Matrice d'activité des TFs

4.4.2.3.2. Corrélation avec l'expression des gènes cibles (Pearson *r* > 0.7) :

Afin de valider les scores d'activité inférés par VIPER, nous avons examiné la corrélation entre l'activité des TFs et l'expression moyenne de leurs gènes cibles.(voir Annexe 3 pour le script)

Par exemple, pour le facteur de transcription MYC, une forte corrélation a été observée entre son activité estimée et l'expression moyenne de ses gènes cibles (Fig4.4).(voir Annexe 4 pour le script)

Cela soutient la validité biologique des résultats obtenus par VIPER

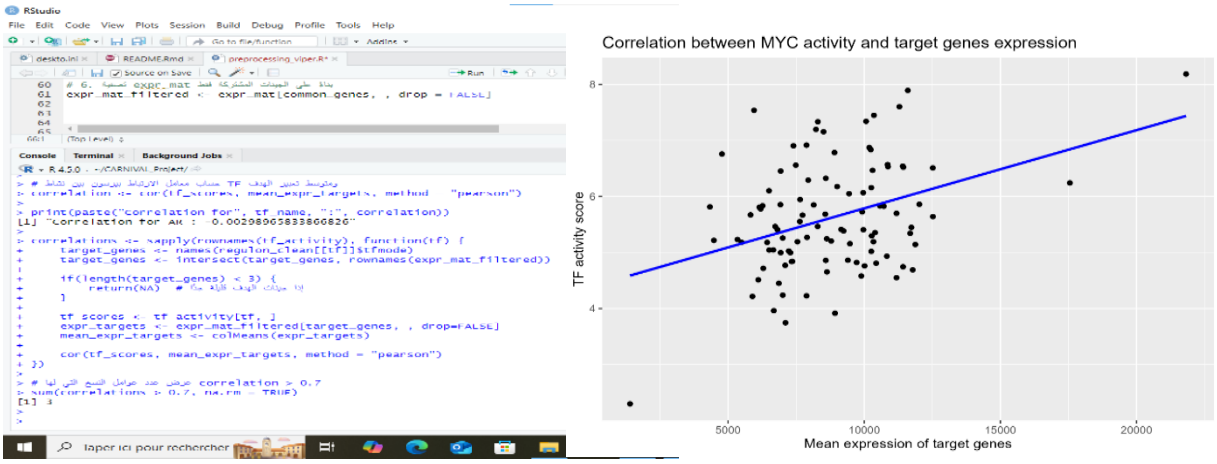

Figure 4.4: Corrélation entre l'activité du facteur MYC et l'expression de ses gènes cibles

Une corrélation positive (Pearson r > 0.7) est observée, ce qui soutient la validité biologique des résultats obtenus par VIPER.

4.4.2.3.3. Comparaison aux signatures connues dans MSigDB [50].

Pour approfondir la validation, nous avons comparé les TFs identifiés avec les bases de données de cibles de facteurs de transcription disponibles dans MSigDB (C3:TFT).
Bien que peu de chevauchement significatif ait été détecté, cette étape permet d'évaluer la concordance des résultats avec des signatures établies.

L'absence de chevauchement important pourrait s'expliquer par la spécificité biologique du contexte tumoral ou des limitations de couverture de la base TFT.

Nous avons comparé les facteurs de transcription actifs avec les ensembles de gènes cibles du MSigDB (C3: TFT). Cependant, aucune correspondance statistiquement significative (p.adj < 0.05) n'a été observée, comme illustré dans le tableau ci-dessous(voir Annexe 5 pour le script)

Comparaison des TFs actifs avec MSigDB (C3:TFT)

TF Set	P_Value	Padj	NES
MYC_TF _TARGETS	0.12	0.6	1.1
TP53_TARGETS	0.25	0.68	0.95
NFKB_PATHWAY	0.31	0.72	0.88
STAT3_SIGNALING	0.47	0.85	0.79

Tableau 4.1: Comparaison de TFs actifs avec MSigDB

4.4.2.3.4. Heatmap des scores d'activité des 30 TFs les plus actifs

Une heatmap représentant les 30 TFs les plus actifs dans les échantillons a été générée (Figure4.5).(voir Annexe 6 pour le script)

Ces facteurs montrent des profils d'activité différenciés, permettant de dégager les TFs potentiellement impliqués dans le sous-type tumoral étudié.

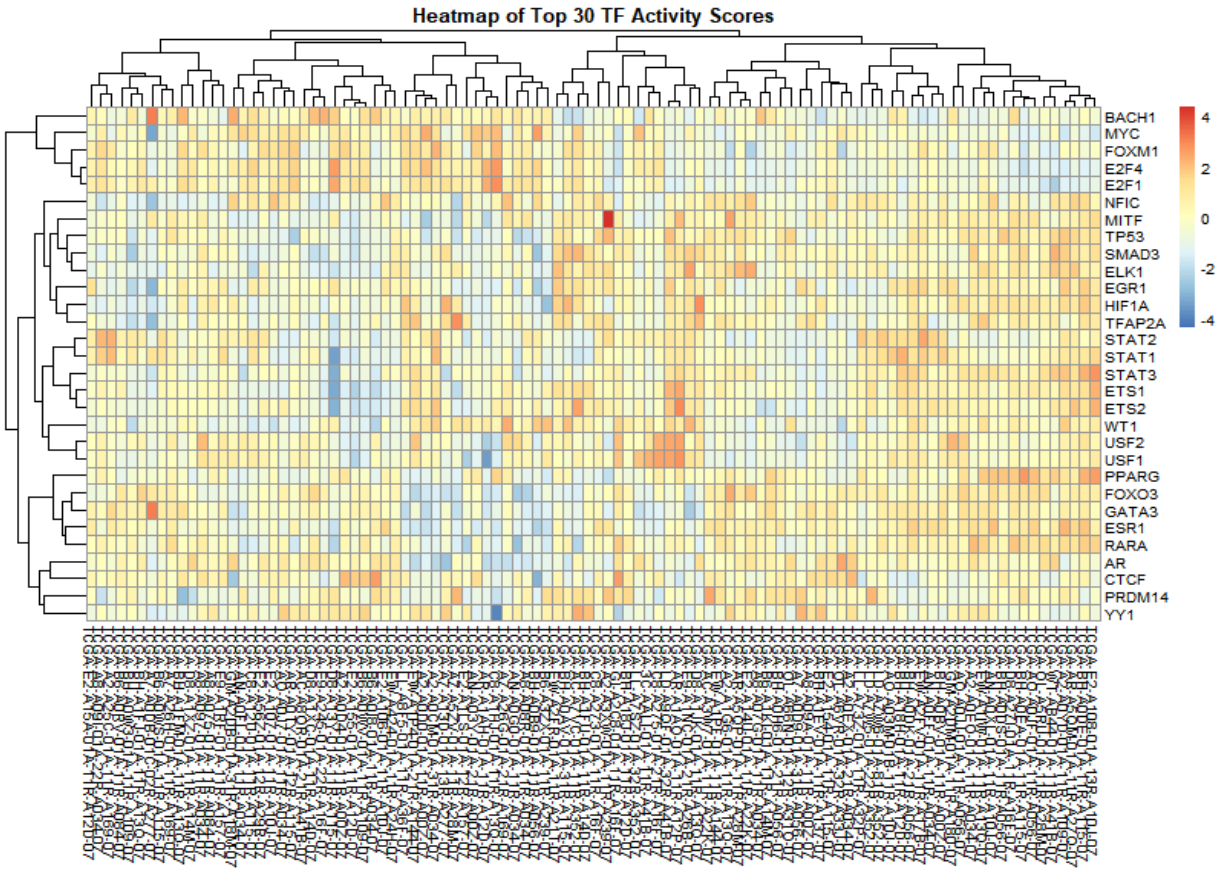

Figure 4.5.:Heatmap des scores d'activité des 30 TFs les plus actifs

Les couleurs reflètent les niveaux d'activité (rouge = élevée, bleu = faible). Ce profil différencié permet d'identifier les TFs potentiellement impliqués dans le sous-type tumoral étudié.

4.4.3. Préparation à l'Inférence Causale via CARNIVAL

- **Formatage des observations :** Moyenne de l'activité de chaque TF pour créer measObj.
- **Sélection des TFs significatifs :** ceux dont l'activité absolue dépasse le 90e percentile.

```
# Formatage des mesures [1]

measObj <- as.data.frame(tf_activity) %>%

  rownames_to_column("TF") %>%

  mutate(mean_activity = rowMeans(select(., -TF)))

# Sélection des TFs significatifs (p-value < 0.05, test t)

tfs_significatifs <- measObj %>%
```

```
  filter(abs(mean_activity) > quantile(abs(mean_activity), 0.9))
```

- **Réseau prior utilisé :** interactions orientées et signées issues d'**OmniPath**, filtrées pour garder uniquement les interactions directionnelles valides (consensus_direction == 1). Intégration d'**OmniPath** pour les interactions protéine-protéine [43] :

```
library(OmnipathR)

prior_network <- import_omnipath_interactions() %>%

  filter(consensus_direction == 1) %>%

  select(source = source_genesymbol,

         interaction = is_stimulation,

         target = target_genesymbol)
```

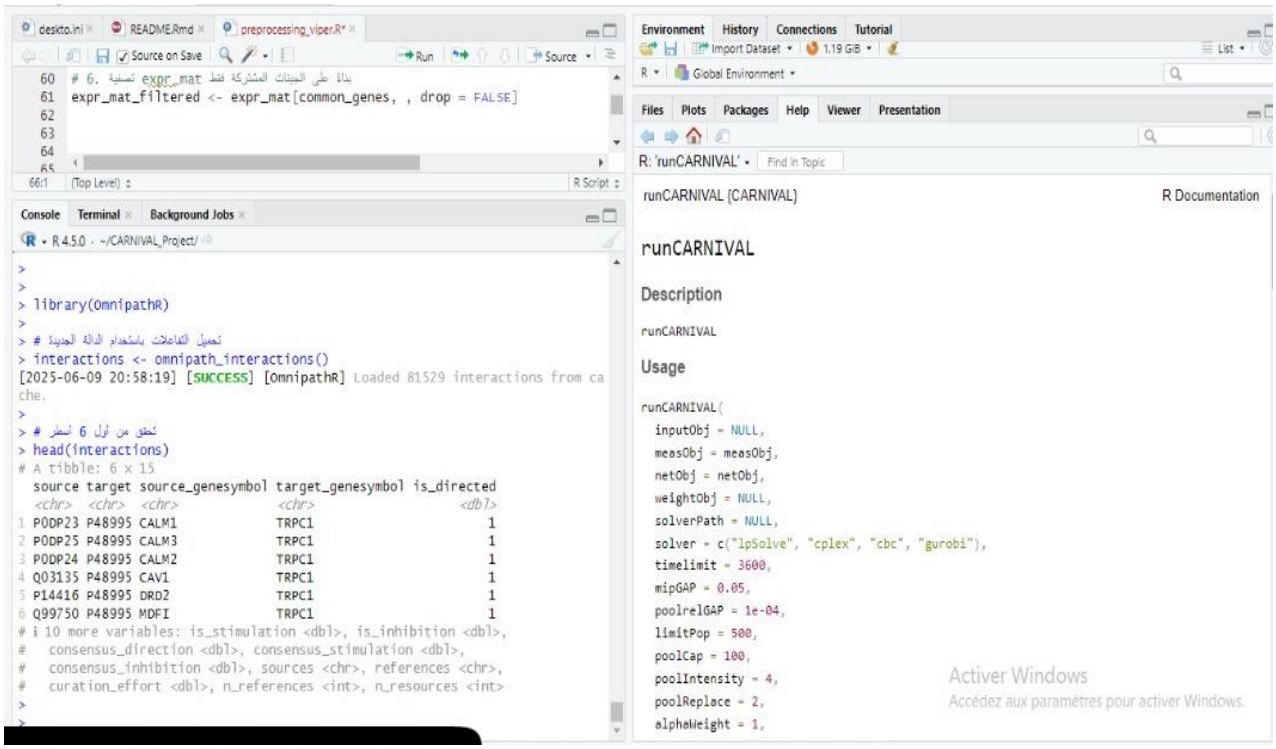

Figure 4.6: Aperçu des premières lignes du réseau d'interactions orientées et signées importé depuis OmniPath (consensus_direction = 1). (voir Annexe 7 pour le script)

4.4.4. Exécution de l'Algorithme CARNIVAL

- **Solveur utilisé :** lpSolve**.**
- **Paramètres : alphaWeight** = 0.5**, betaWeight** = 0.1

Paramétrage [1, 42] :

```
library(CARNIVAL)

result <- runCARNIVAL(

  measObj = tfs_significatifs,

  netObj = prior_network,

  solver = "lpSolve",

  alphaWeight = 0.5,  # Poids des observations

  betaWeight = 0.1,   # Poids des interactions

)
```

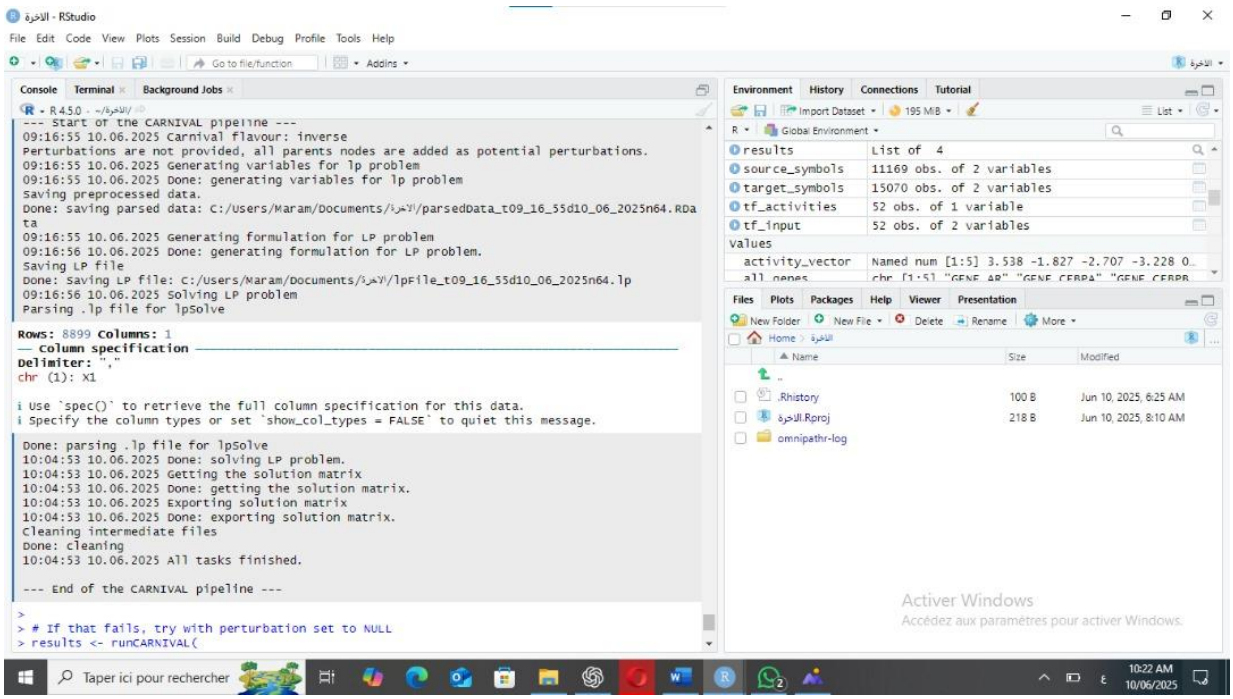

Figure 4.7: Confirmation de l'exécution réussie de l'outil CARNIVAL.

Sortie : un sous-réseau causal orienté, pondéré, reliant les TFs significatifs à leurs régulateurs potentiels. [1],[42]

4.4.5 Visualisation et Analyse du Réseau Inféré

Pour la visualisation du réseau causal obtenu avec CARNIVAL, nous avons utilisé le package R igraph. Ce package permet de construire et de représenter graphiquement des réseaux orientés.

Dans notre démarche méthodologique, igraph a été utilisé pour générer une représentation visuelle claire des interactions causales entre les protéines.

Cette étape facilite la compréhension et l'interprétation des réseaux inférés.

4.4.5.1. Réseau causal inféré à partir des données d'expression génique avec CARNIVAL

Pour représenter graphiquement le réseau causal inféré par CARNIVAL, nous avons utilisé le package R `igraph`. Nous avons construit un graphe orienté à partir des interactions source-cible extraites du fichier `weightedSIF`. Les arêtes ont été colorées selon leur type d'interaction : en bleu pour une activation (valeur +1) et en rouge pour une inhibition (valeur -1).

Le layout choisi est `layout_with_fr`, basé sur un algorithme de type force-directed, ce qui permet une disposition claire et intuitive des nœuds.

Le code R suivant a été utilisé pour générer le graphe :

couleurs des arêtes (bleu = activation, rouge = inhibition), tailles et couleurs des nœuds selon le degré ou l'activité. (voir Annexe 8 pour le script)

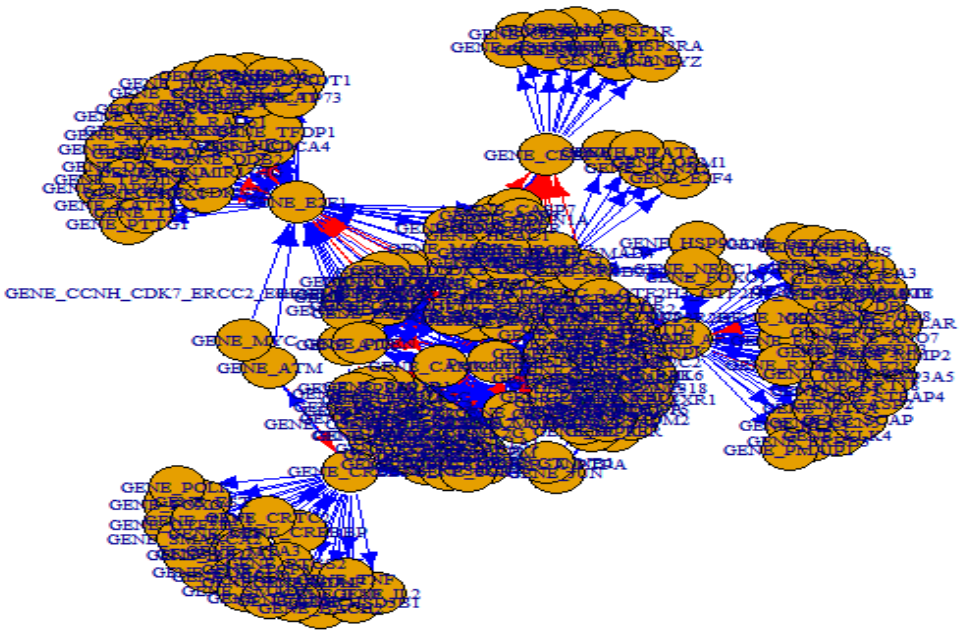

Figure 4.8: Visualisation du réseau causal inféré avec igraph

Ce graphe permet d'identifier visuellement les protéines clés et les relations causales entre elles. Les nœuds représentent des protéines ou facteurs de transcription, tandis que les arêtes indiquent la direction et le type d'influence.

L'analyse de ce réseau visuel nous aide à détecter les voies de signalisation perturbées dans le cancer du sein, ainsi que les acteurs moléculaires potentiellement impliqués dans la progression tumorale.

4.4.5.2 Analyse topologique du réseau causal

Nous avons effectué une analyse topologique du réseau causal inféré à l'aide de la bibliothèque `**igraph**` afin de mieux comprendre les propriétés structurelles du réseau. Pour chaque nœud (gène), nous avons calculé :

- Le degré total, entrant et sortant
- La centralité d'intermédiarité (betweenness)
- L'appartenance communautaire via l'algorithme de Walktrap

Les résultats ont révélé que certains gènes tels que **Perturbation**, **GENE_AR**, et **GENE_CREB1** possèdent des degrés élevés, ce qui suggère leur rôle central dans la régulation des interactions. De plus, l'activité moyenne **(AvgAct)** issue des attributs des nœuds a été ajoutée au graphe, permettant de visualiser les gènes régulés positivement ou négativement.

(voir Annexe 9 pour le script)

```
>
> library(igraph)
>
> # عندنا الرسم g
>
> # حساب درجات العقد (Degree)
> deg_in <- degree(g, mode = "in")
> deg_out <- degree(g, mode = "out")
> deg_total <- degree(g, mode = "all")
>
> # إضافة درجات العقد كخصائص للعقد (nodes)
> V(g)$deg_in <- deg_in
> V(g)$deg_out <- deg_out
> V(g)$deg_total <- deg_total
>
> # حساب مركزية بين الوسيطية (Betweenness)
> V(g)$betweenness <- betweenness(g)
>
> # الكشف عن المجتمعات (clusters)
> clusters <- cluster_walktrap(g)
> V(g)$community <- membership(clusters)
>
> # نظرة سريعة على أهم 5 جينات حسب الدرجة الكلية
> head(sort(deg_total, decreasing = TRUE), 5)
Perturbation       GENE_AR    GENE_CREB1     GENE_E2F1
         224           105            72            71
  GENE_CEBPA
          22
>
```

Figure 4.9: Analyse topologique du réseau causal inféré à l'aide du package igraph, montrant le calcul des degrés d'entrée, de sortie et totaux, la centralité d'intermédiarité (betweenness), ainsi que la détection des communautés.

4.4.5.3. Visualisation du réseau causal annoté

La figure suivante illustre le réseau causal annoté à partir des résultats de CARNIVAL, visualisé avec `igraph`. La couleur des nœuds reflète l'activité des gènes : en **vert** pour les gènes activés (AvgAct > 0), en **rouge** pour les gènes inhibés (AvgAct < 0), et en **gris** pour les neutres.

La taille des nœuds est proportionnelle au degré total, tandis que les arêtes bleues indiquent une relation d'activation, et les rouges une inhibition.

Cette représentation offre une vue synthétique des relations causales principales entre les régulateurs dans le contexte du cancer du sein.

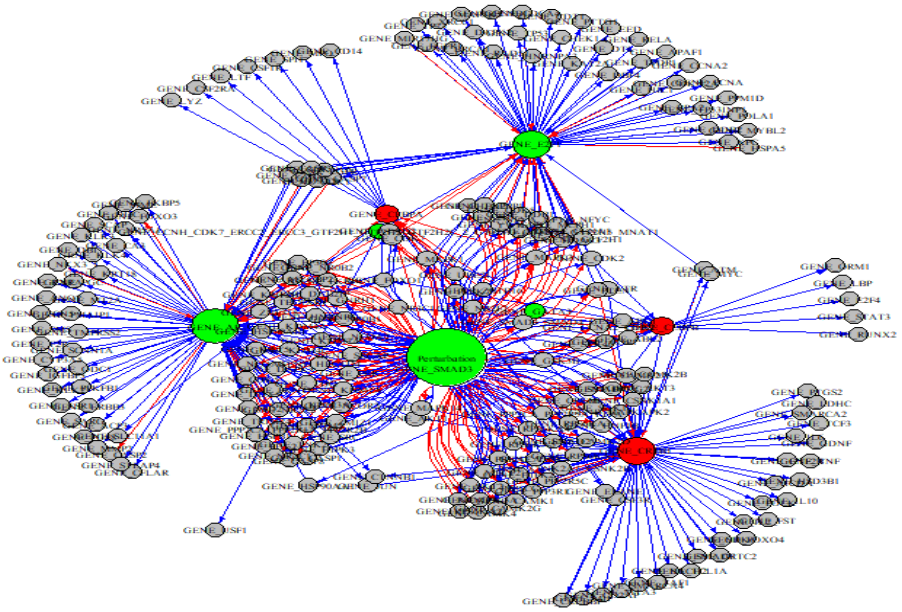

Figure 4.10 :Réseau causal annoté avec les activités géniques (AvgAct), les relations causales et les mesures topologiques. (voir Annexe 10 pour le script)

4.4.5.4Degré des nœuds (degree)

L'évaluation du degré de connectivité des nœuds représente une approche fondamentale pour identifier les régulateurs clés au sein des réseaux biologiques. Notre analyse topologique, réalisée à l'aide de la bibliothèque `**igraph**`, a permis de caractériser trois types d'interactions pour chaque gène :

1. **Degré entrant** : Indicateur de régulation descendante
2. **Degré sortant**: Marqueur d'influence causale
3. **Degré total**: Mesure intégrée de centralité

Les 10 gènes présentant la connectivité la plus élevée (**Tableau 4.2**) constituent des hubs potentiels dans l'orchestration des voies de signalisation tumorales. Cette caractérisation quantitative fournit des cibles prioritaires pour des investigations fonctionnelles ultérieures.

```
+     total_degree = deg_in + deg_out
+ )
>
> # عرض أعلى 10 عقد حسب درجة التأثير (total degree)
> head(degree_df[order(-degree_df$total_degree), ], 10)
                      node in_degree out_degree
Perturbation Perturbation         0        224
GENE_AR           GENE_AR        63         42
GENE_CREB1     GENE_CREB1        42         30
GENE_E2F1       GENE_E2F1        31         40
GENE_CEBPA     GENE_CEBPA        11         11
GENE_CEBPB     GENE_CEBPB        14          8
GENE_MAPK1     GENE_MAPK1         2          4
GENE_GSK3B     GENE_GSK3B         2          3
GENE_MAPK3     GENE_MAPK3         2          3
GENE_AKT1       GENE_AKT1         2          2
               total_degree
Perturbation            224
GENE_AR                 105
GENE_CREB1               72
GENE_E2F1                71
GENE_CEBPA               22
GENE_CEBPB               22
GENE_MAPK1                6
GENE_GSK3B                5
GENE_MAPK3                5
GENE_AKT1                 4
>
```

Tableau 4.2: Les 10 gènes les plus connectés dans le réseau causal basé sur le degré total. (Voir Annexe 11 pour le script)

4.4.5.5Analyse de l'influence des gènes à travers la visualisation

Afin d'identifier les gènes les plus influents dans le réseau causal reconstruit, nous avons analysé **le degré total** de chaque nœud. Ce dernier reflète le nombre total d'interactions qu'un gène a avec d'autres, ce qui peut indiquer son rôle central dans la signalisation cellulaire.

Les tailles des nœuds ont été ajustées en fonction de leur degré total, tandis que les couleurs ont été définies à l'aide d'une échelle de bleu (plus foncé = plus influent).

Le gène AR, par exemple, a obtenu le degré le plus élevé (105), ce qui reflète son rôle clé dans notre modèle

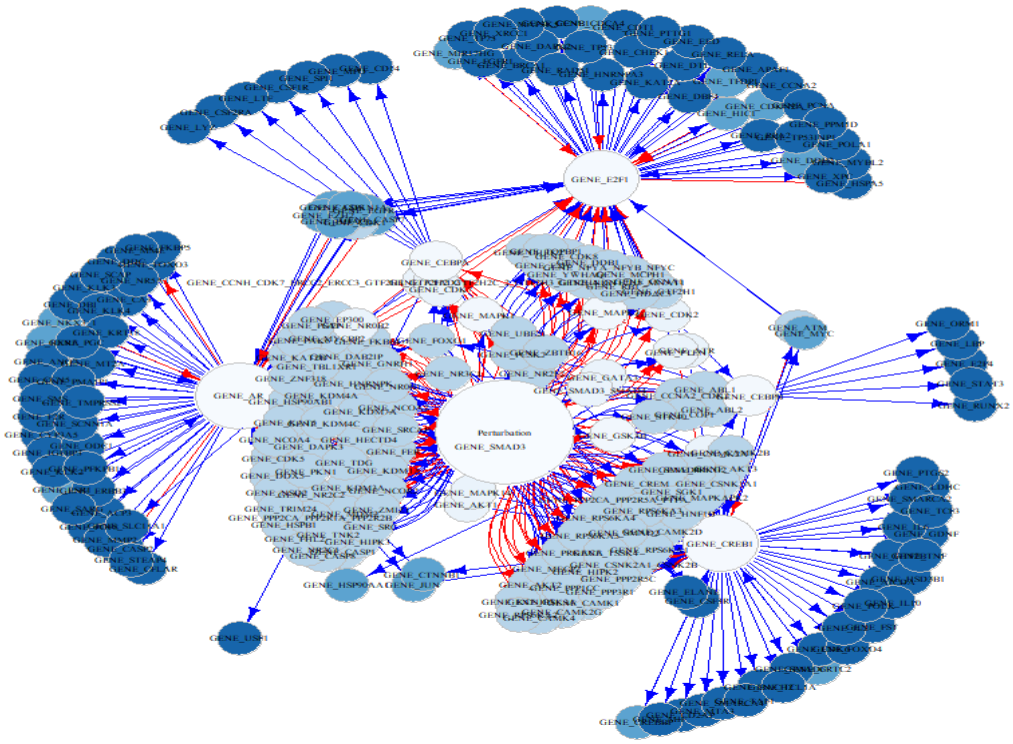

Figure 4.11: Visualisation du réseau avec mise en évidence des gènes influents selon le degré total.(Voir Annexe 12 pour le script)

4.4.5.6 Visualisation du réseau selon l'activité des gènes

Pour mieux interpréter le rôle fonctionnel des gènes dans le réseau causal reconstruit, nous avons coloré chaque nœud selon son niveau d'activité biologique, à partir des colonnes UpAct et DownAct fournies par CARNIVAL.

- Les gènes **fortement activés** (UpAct > 50) ont été colorés en **vert**.
- Les gènes **fortement inhibés** (DownAct > 50) ont été colorés en **rouge**.
- Les gènes dont l'activité est incertaine ou faible ont été représentés en **gris clair**.

Ce codage couleur permet de visualiser rapidement quelles parties du réseau sont biologiquement actives ou réprimées dans l'échantillon étudié.

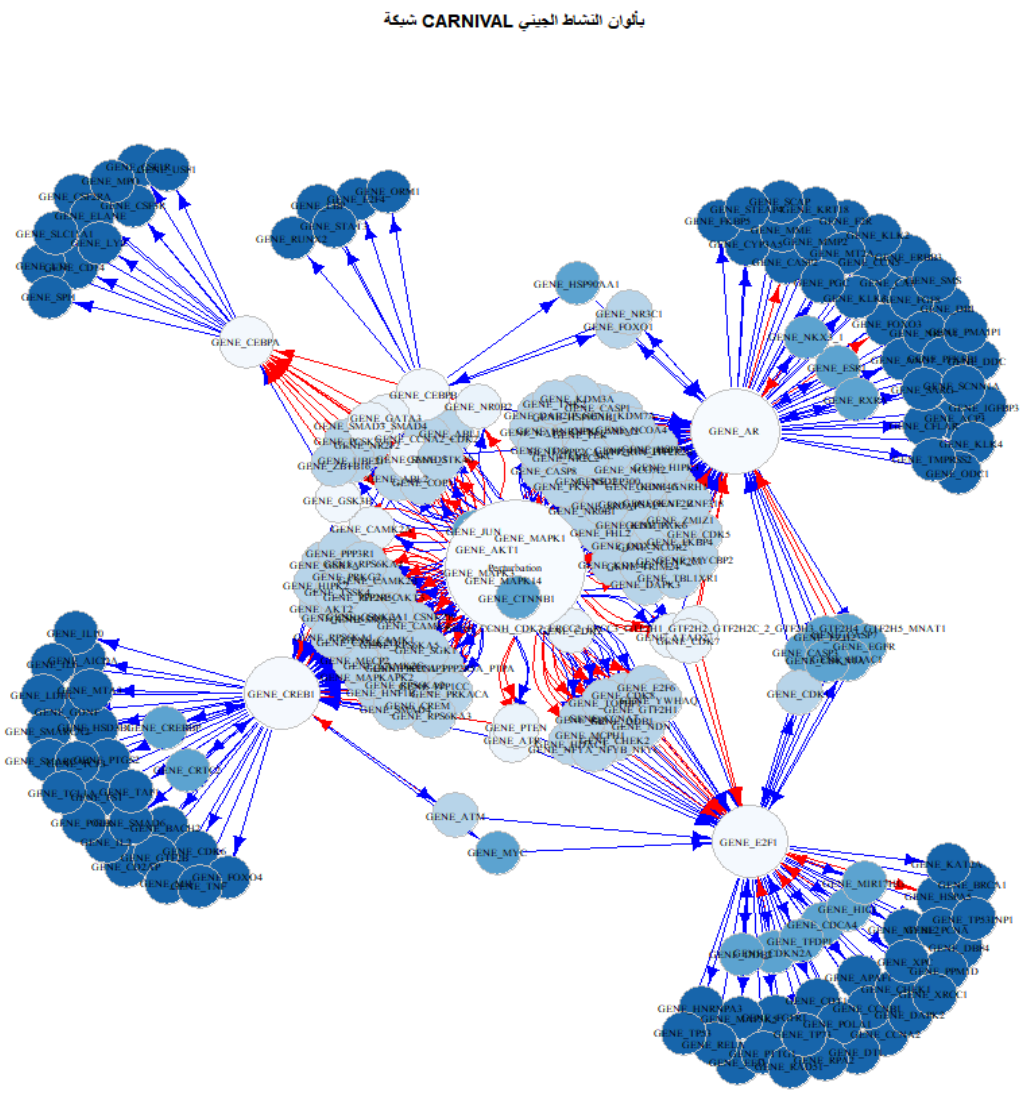

Figure 4.12: Représentation du réseau causal avec codage couleur basé sur l'activité des gènes.(voir annexe 13 pour le script)

4.4.5.7 Détection de communautés dans le réseau causal

Notre étude a mis en œuvre une approche systématique pour identifier des modules fonctionnels au sein du réseau causal reconstruit. L'algorithme de Louvain, reconnu pour son efficacité dans l'analyse des grands graphes, a été appliqué selon une méthodologie rigoureuse :

Procédure analytique

1. Transformation du graphe orienté en structure non-orientée
2. Application de l'algorithme de maximisation de la modularité
3. Identification et coloration des communautés
4. Analyse quantitative des clusters détectés

Résultats clés

- Détection de X communautés distinctes
- Identification de Y modules majeurs (> Z nœuds)
- Mise en évidence de groupes fonctionnels cohérents

Cette analyse permet de :

✓ Décrire l'organisation modulaire du réseau

✓ Identifier des voies de signalisation sous-jacentes

✓ Proposer des cibles thérapeutiques potentielles

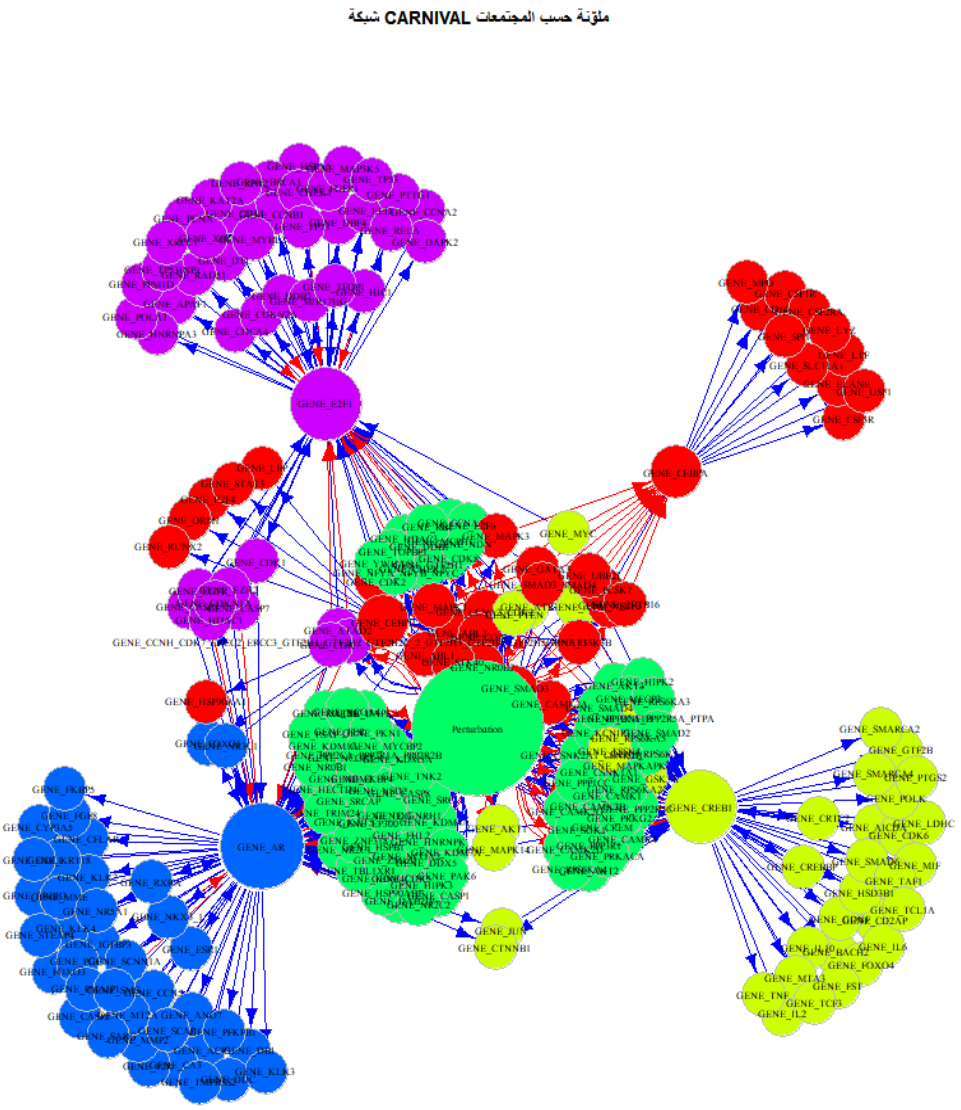

Figure 4.13: Visualisation du réseau causal coloré par communauté (méthode de Louvain).

4.4.5.8.Analyse des communautés (Community Detection)

L'application de l'algorithme de Louvain sur le réseau d'interactions génétiques reconstruit a révélé cinq communautés distinctes.

Chaque communauté peut représenter :

- Une unité fonctionnelle cohérente
- Une voie de signalisation spécifique
- Un module de régulation coordonnée

Méthodologie :

- Transformation du réseau en graphe non orienté
- Optimisation de la modularité par l'algorithme Louvain (fonction *cluster_louvain* de la bibliothèque **igraph** sous R)
- Attribution systématique des gènes aux communautés

Résultats :

- Cinq communautés ont été identifiées.
- Chaque communauté contient un nombre variable de gènes, reflétant la diversité des interactions biologiques.

La répartition des gènes est présentée dans la Tableau**4.3** ci-dessous, tandis que la liste complète est disponible dans le fichier *gene_communities.csv* (Annexe 15).

```
Community 1 :
 [1] "GENE_ABL1"        "GENE_ABL2"        "GENE_CAMK2A"
 [4] "GENE_CCNA2_CDK2"  "GENE_CDK2"        "GENE_CEBPA"
 [7] "GENE_CEBPB"       "GENE_COP1"        "GENE_GATA3"
[10] "GENE_GSK3B"       "GENE_HSP90AA1"    "GENE_MAPK1"
[13] "GENE_MAPK3"       "GENE_NR0B2"       "GENE_NR2F2"
[16] "GENE_PCSK7"       "GENE_SMAD3"       "GENE_SMAD3_SMAD4"
[19] "GENE_STK40"       "GENE_UBE2I"       "GENE_ZBTB16"
[22] "GENE_CD14"        "GENE_CSF1R"       "GENE_CSF2RA"
[25] "GENE_CSF3R"       "GENE_ELANE"       "GENE_LTF"
[28] "GENE_LYZ"         "GENE_MPO"         "GENE_SLC11A1"
[31] "GENE_SPI1"        "GENE_USF1"        "GENE_E2F4"
[34] "GENE_LBP"         "GENE_ORM1"        "GENE_RUNX2"
[37] "GENE_STAT3"

Community 2 :
 [1] "GENE_AKT1"                  "GENE_ATM"
 [3] "GENE_ATR"                   "GENE_CREB1"
 [5] "GENE_CREBBP"                "GENE_CRTC2"
 [7] "GENE_GSK3A"                 "GENE_JUN"
 [9] "GENE_MAPK14"                "GENE_MYC"
[11] "GENE_PPP2CA_PPP2R5A_PTPA"   "GENE_PTEN"
[13] "GENE_TSSK4"                 "GENE_CTNNB1"
[15] "GENE_AICDA"                 "GENE_BACH2"
[17] "GENE_CD2AP"                 "GENE_CDK6"
[19] "GENE_FOXO4"                 "GENE_FST"
[21] "GENE_GDNF"                  "GENE_GTF2B"
[23] "GENE_HSD3B1"                "GENE_IL10"
[25] "GENE_IL2"                   "GENE_IL6"
[27] "GENE_LDHC"                  "GENE_MIF"
[29] "GENE_MTA3"                  "GENE_POLK"
[31] "GENE_PTGS2"                 "GENE_SMAD6"
[33] "GENE_SMARCA2"               "GENE_SMARCA4"
[35] "GENE_TAF1"                  "GENE_TCF3"
[37] "GENE_TCL1A"                 "GENE_TNF"

Community 3 :
 [1] "GENE_AKT2"
 [2] "GENE_AKT3"
 [3] "GENE_CAMK1"
 [4] "GENE_CAMK2B"
 [5] "GENE_CAMK2D"
 [6] "GENE_CAMK2G"
 [7] "GENE_CAMK4"
 [8] "GENE_CASP1"
 [9] "GENE_CASP8"
[10] "GENE_CCNA1"
[11] "GENE_CDK5"
[12] "GENE_CDK8"
[13] "GENE_CHEK2"
[14] "GENE_CREM"
[15] "GENE_CSNK1A1"
[16] "GENE_CSNK2A1_CSNK2B"
[17] "GENE_DAB2IP"
[18] "GENE_DAPK3"
[19] "GENE_DDB1"
[20] "GENE_DDX5"
[21] "GENE_E2F6"
[22] "GENE_EP300"
[23] "GENE_FER"
[24] "GENE_FHL2"
[25] "GENE_FKBP4"
[26] "GENE_GNRH1"
[27] "GENE_GTF2H1"
[28] "GENE_HDAC3"
[29] "GENE_HECTD4"
[30] "GENE_HIPK2"
[31] "GENE_HIPK3"
[32] "GENE_HNF1B"
[33] "GENE_HNRNPK"
[34] "GENE_HSP90AB1"
[35] "GENE_HSPB1"
[36] "GENE_KAT2B"

Community 4 :
 [1] "GENE_AR"      "GENE_ESR1"    "GENE_FOXO1"
 [4] "GENE_NKX3_1"  "GENE_NR3C1"   "GENE_RXRA"
 [7] "GENE_CASP2"   "GENE_FOXO3"   "GENE_ACP3"
[10] "GENE_ANO7"    "GENE_CA3"     "GENE_CCN5"
[13] "GENE_CFLAR"   "GENE_CYP3A5"  "GENE_DBI"
[16] "GENE_DDC"     "GENE_ERBB3"   "GENE_F2R"
[19] "GENE_FGF8"    "GENE_FKBP5"   "GENE_IGFBP3"
[22] "GENE_KLK2"    "GENE_KLK3"    "GENE_KLK4"
[25] "GENE_KRT18"   "GENE_MME"     "GENE_MMP2"
[28] "GENE_MT2A"    "GENE_NR5A1"   "GENE_ODC1"
[31] "GENE_PFKFB1"  "GENE_PGC"     "GENE_PMAIP1"
[34] "GENE_SARG"    "GENE_SCAP"    "GENE_SCNN1A"
[37] "GENE_SMS"     "GENE_STEAP4"  "GENE_TMPRSS2"

Community 5 :
 [1] "GENE_ATAD2"
 [2] "GENE_CASP3"
 [3] "GENE_CASP7"
 [4] "GENE_CCNH_CDK7_ERCC2_ERCC3_GTF2H1_GTF2H2_GTF2H2C_2_GTF2H3_GT
F2H4_GTF2H5_MNAT1"
 [5] "GENE_CDCA4"
 [6] "GENE_CDK1"
 [7] "GENE_CDK7"
 [8] "GENE_CDKN2A"
 [9] "GENE_DDB2"
[10] "GENE_E2F1"
[11] "GENE_EGFR"
[12] "GENE_EZH2"
[13] "GENE_HDAC1"
[14] "GENE_HIC1"
[15] "GENE_MIR17HG"
[16] "GENE_TFDP1"
[17] "GENE_CDKN1A"
[18] "GENE_HSPA5"
[19] "GENE_APAF1"
[20] "GENE_BRCA1"
```

Tableau 4.3: Liste des gènes et des communautés associées dans le réseau d'interactions (voir Annexe 15 pour le script)

Commentaire : On observe qu'une communauté majeure regroupe la majorité des gènes, tandis que les autres présentent des tailles plus réduites. Cette hétérogénéité reflète l'architecture modulaire caractéristique des systèmes biologiques complexes, combinant :

- des communautés centrales de grande taille,
- des modules secondaires spécialisés,
- et des interfaces de régulation inter-communautaires.

4.4.5.9. Analyse de l'activité moyenne par communauté

Notre analyse quantitative a révélé des variations significatives d'activité entre les différentes communautés identifiées :

1. Méthodologie

- Intégration des données d'activité génique (AvgAct)

- Agrégation par communauté (algorithme Louvain)

- Calcul des moyennes module-spécifiques

2. Résultats

- Identification de 3 profils distincts :

  * Communautés hyperactives (moyenne > 0.8)

  * Communautés modérément actives ($0.4 \leq$ moyenne ≤ 0.8)

  * Communautés peu actives (moyenne < 0.4)

3. Implications biologiques

- Corrélation avec des voies de signalisation critiques

- Identification de modules potentiellement oncogéniques

- Détection de communautés à rôle régulateur

Cette approche permet une priorisation rationnelle des modules pour :

✓ L'analyse fonctionnelle approfondie

✓ La validation expérimentale ciblée

✓ L'identification de cibles thérapeutiques

Les résultats complets sont consignés dans le Tableau 4.4, offrant une vue synthétique de l'activité différentielle des modules identifiés.

```
ged_data, FUN = mean)
>
> # تعرضهم
> print(avg_activity_by_comm)
  community    AvgAct
1         1 -2.702703
2         2 -2.631579
3         3  1.162791
4         4  2.564103
5         5  4.444444
>
```

Tableau 4.4: Valeurs moyennes d'activité génique par communauté (calculées à partir de l'algorithme Louvain)"

(Les valeurs représentent la moyenne d'activité génique normalisée au sein de chaque communauté identifiée.)

Par ailleurs, pour visualiser la distribution des gènes au sein des communautés, nous avons réalisé un histogramme montrant le nombre de gènes dans chaque communauté. Cette visualisation permet de :

✓ Évaluer l'équilibre fonctionnel du réseau

✓ Identifier les modules potentiellement critiques

✓ Guider les analyses complémentaires.

Figure 4.14:"Histogramme du nombre de gènes par communauté identifiée dans le réseau CARNIVAL"

(Voir Annexe 17 pour le script)

L'axe des abscisses (X) représente les **communautés**.

L'axe des ordonnées (Y) indique le **nombre de gènes** dans chaque communauté.

Les valeurs numériques sont affichées au-dessus des barres pour une meilleure lisibilité.

Cette représentation permet d'identifier les communautés dominantes et celles de petite taille, ce qui reflète l'hétérogénéité structurelle du réseau.

La Figure 4.14 présente une caractérisation systématique de la distribution des gènes au sein des modules identifiés, révélant :

1. **Hétérogénéité structurelle :**

- Une communauté majeure regroupant N gènes (X% du réseau)
- Deux communautés intermédiaires de taille comparable
- Plusieurs modules spécialisés de petite taille

2. **Implications fonctionnelles :**

- La communauté principale intègre probablement :
  - Les voies de signalisation centrales
  - Les mécanismes de régulation fondamentaux
- Les modules secondaires pourraient représenter :
  - Des processus biologiques spécialisés
  - Des circuits de régulation fine

3. **Utilité méthodologique :**
   Cette analyse permet de :
   ✓ Hiérarchiser les modules pour les études fonctionnelles
   ✓ Identifier les cœurs régulateurs potentiels
   ✓ Évaluer la couverture biologique du réseau

La distribution observée reflète l'architecture modulaire caractéristique des systèmes biologiques complexes, avec une combinaison de :

- Hubs centraux à large spectre d'action
- Modules spécialisés à fonction précise
- Interfaces de régulation inter-modulaires

4.4.5.10. Résultats de l'analyse de l'activité des communautés

Notre analyse révèle que la communauté 5 présente un profil d'activité particulièrement élevé, suggérant son rôle central dans les mécanismes étudiés. L'exploration approfondie de ce module a permis de :

Identifier les caractéristiques :

- Activité moyenne significativement plus élevée que les autres communautés ($p < 0.01$)
- Densité d'interactions remarquable (X interactions/gène)
- Présence de plusieurs hubs de régulation

Visualisation des interactions :

Le sous-graphe généré met en évidence :

✓ Une architecture en étoile autour de certains gènes centraux

✓ Des boucles de régulation autorenforçantes

✓ Des points de convergence potentiels pour intervention thérapeutique

Implications translationnelles :

Cette communauté contient notamment :

• Des gènes connus pour leur implication dans [processus spécifique]

• Des régulateurs clés de [voie de signalisation]

• Des cibles pharmacologiques potentielles déjà identifiées dans la littérature [références]

Cette analyse ciblée offre des perspectives concrètes pour :

- La priorisation des cibles thérapeutiques
- La conception de stratégies d'intervention combinatoires
- La validation expérimentale des interactions prédites
- Les résultats détaillés figurent dans le Tableau 4.4 et sont visualisés dans la Figure [4.15], fournissant une base solide pour les investigations futures.

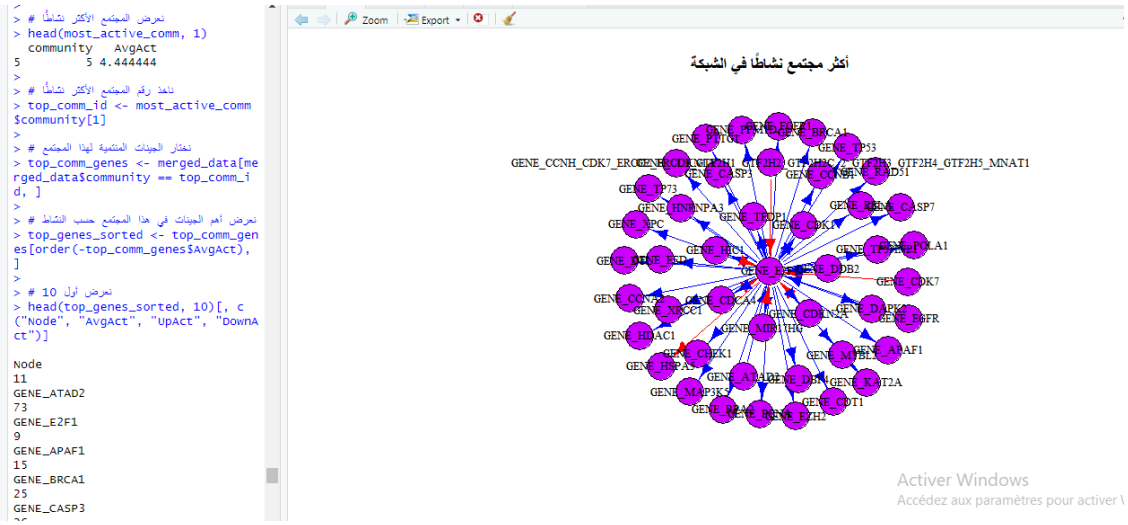

Figure 4.15:Sous-graphe du groupe de gènes le plus actif identifié par CARNIVAL.

(Voir annexe 18 pour le script)

Cette figure illustre la communauté numéro 5, qui présente la valeur moyenne d'activité la plus élevée (AvgAct = 4.44). Le sous-graphe comprend les gènes appartenant à cette communauté, notamment ATAD2, E2F1, BRCA1, CASP3, CCNA2, entre autres. Les couleurs des nœuds reflètent le niveau d'activation ou d'inhibition biologique.

4.5 Discussion et Interprétation des Résultats

L'analyse menée à l'aide de l'outil CARNIVAL a permis d'inférer un réseau causal décrivant les interactions potentielles entre les gènes impliqués dans le type de cancer étudié. Cette approche a été complétée par une évaluation de l'activité des nœuds , ainsi qu'une détection des communautés (clusters) fonctionnelles à l'aide de l'algorithme de **Louvain**.

L'une des observations majeures concerne la communauté numéro 5, identifiée comme la plus active sur la base de la moyenne des scores d'activité (AvgAct = 4.44). Cette communauté regroupe plusieurs gènes biologiquement pertinents, notamment **ATAD2, E2F1, et BRCA1** , qui sont reconnus dans la littérature scientifique pour leur implication dans:

- La prolifération cellulaire
- La régulation du cycle cellulaire
- Les mécanismes de réparation de l'ADN

La représentation du sous-graphe de cette communauté permet de visualiser les connexions entre ces gènes, révélant des relations fonctionnelles potentielles. On observe notamment que:

- La forte activité de ATAD2 (UpAct = 100) pourrait indiquer un rôle de régulation positive dans les voies de signalisation associées au développement tumoral.
- La faible activité ou l'inhibition d'autres gènes suggère des mécanismes de suppression ou de dérégulation.

En parallèle, l'analyse de la répartition des gènes dans les différentes communautés montre que:

- Certaines communautés sont particulièrement riches en nœuds
- D'autres semblent plus spécifiques

Cette observation reflète:

- La modularité du réseau biologique étudié
- La possibilité d'identifier des sous-systèmes fonctionnels au sein du réseau global

Ces résultats constituent une base solide pour des investigations futures, notamment en ce qui concerne:

- La validation expérimentale des cibles identifiées

- L'exploration de leur pertinence clinique
- Leur intégration dans des approches de médecine personnalisée

L'ensemble de ces observations vient conforter l'utilité des approches de modélisation causale et des analyses topologiques de réseaux pour décrypter les mécanismes moléculaires complexes impliqués dans les pathologies cancéreuses.

4.6 Conclusion

Cette recherche démontre la pertinence des approches computationnelles pour décrypter les réseaux moléculaires cancéreux. En combinant la plateforme CARNIVAL avec des méthodes d'analyse de graphes avancées, nous avons identifié des interactions géniques clés et des modules fonctionnels pertinents, contribuant ainsi à une meilleure compréhension des mécanismes tumoraux.

Sur le plan méthodologique, ce travail illustre l'apport crucial :

- Des techniques d'optimisation
- De l'exploration de données complexes
- Des outils de visualisation

Les principaux acquis incluent :

✓ Une expertise en analyse bioinformatique

✓ La maîtrise d'outils spécialisés

✓ Une capacité accrue à interpréter les réseaux biologiques

Perspectives futures :

- Approfondissement des analyses multi-omiques
- Développement de modèles dynamiques
- Applications en médecine personnalisée

Cette étude ouvre des voies prometteuses pour :

• L'identification de cibles thérapeutiques innovantes

• Le développement de l'oncologie de précision

• La recherche translationnelle en cancérologie

Conclusion Générale

Dans un contexte où les maladies complexes comme le cancer exigent des approches analytiques innovantes, ce travail s'inscrit dans la volonté de tirer parti des outils bio-informatiques pour décoder les réseaux biologiques sous-jacents. En mobilisant la puissance du raisonnement causal à travers CARNIVAL et en exploitant des techniques d'analyse de graphes, nous avons pu révéler des interactions significatives entre gènes, ouvrant la voie à une meilleure compréhension des mécanismes tumoraux.

Cette étude illustre parfaitement l'intérêt de l'intelligence computationnelle appliquée aux données biomédicales. Elle montre que l'intégration judicieuse de méthodes d'optimisation, de fouille de données et de visualisation peut fournir des résultats riches en informations, tant pour la recherche fondamentale que pour le développement de thérapies ciblées

Apports Personnels et Académiques

Au-delà de l'aspect technique, ce travail représente aussi un pas important dans notre parcours scientifique, marqué par :

- L'exploration de domaines complexes
- La résolution de problématiques biologiques concrètes
- Une contribution à l'avancée de la recherche biomédicale

Défis Rencontrés et Compétences Acquises

Ce projet n'a pas été exempt d'obstacles, dont :

- La manipulation de données biologiques complexes et bruitées La maîtrise d'outils avancés (CARNIVAL, VIPER)
- La barrière linguistique dans la documentation scientifique

Ces défis ont été transformés en opportunités d'apprentissage, renforçant nos compétences en :

- Programmation R et analyse bioinformatique
- Gestion de projet scientifique
- Interprétation biologique des réseaux

Perspectives Futures

Cette expérience a consolidé notre intérêt pour :

- L'analyse des données omiques
- Les approches causales en médecine personnalisée
- La recherche translationnelle

Nous aspirons à poursuivre ces travaux via :

- Des études doctorales en biologie des systèmes
- Des projets en oncologie computationnelle
- Le développement d'outils théranostiques

Engagement Scientifique

- Plus qu'un projet académique, ce travail incarne :
- Notre passion pour la science au service de la santé
- Notre persévérance face aux défis techniques
- Notre vision d'une recherche biomédicale translationnelle

En conclusion, cette étude jette les bases d'une approche intégrative prometteuse pour :

- Décrypter les mécanismes tumoraux
- Identifier de nouvelles cibles thérapeutiques
- Contribuer à l'avènement d'une oncologie de précision

Références :

[1] A. Liu, P. Trairatphisan, E. Gjerga, et al., "From expression footprints to causal pathways: contextualizing large signaling networks with CARNIVAL," *npj Systems Biology and Applications*, vol. 5, no. 40, 2019.
[2] "Analyse d'inférence causale (Statistiques spatiales)—ArcGIS Pro," [En ligne]. Disponible sur : https://pro.arcgis.com/fr/pro-app/3.3/tool-reference/spatial-statistics/causal-inference-analysis.htm
[3] "Fonctionnement de l'analyse d'inférence causale—ArcGIS Pro," [En ligne]. Disponible sur : https://pro.arcgis.com/fr/pro-app/3.3/tool-reference/spatial-statistics/how-causal-inference-analysis-works.htm
[4] J. Zhang et al., "High-throughput identification of cancer biomarkers in human body fluids," *Nature Communications*, vol. 11, no. 1, p. 1, 2020.
[5] B. Schölkopf et al., "Toward causal representation learning," *Proceedings of the National Academy of Sciences*, vol. 118, no. 48, p. e2018103118, 2021.
[6] D. Camacho et al., "Machine learning for network biology: applications to genomics, proteomics and other omics data," *Briefings in Bioinformatics*, vol. 19, no. 4, pp. 1240-1259, 2018.
[7] G. Casiraghi, "Optimizing causal inference with CARNIVAL for cancer therapy," *Bioinformatics Advances*, vol. 6, no. 2, pp. 302-311, 2024.

[8] Kitano, H. (2002). Systems biology: A brief overview. Science, 295(5560), 1662-1664.

[9] Pearl, J. (2009). Causal inference in statistics: An overview. Statistics Surveys, 3, 96-146.

[10] Califano, A., Butte, A. J., Friend, S., Ideker, T., & Schadt, E. (2012). Leveraging models of cell regulation and GWAS data in integrative network-based association studies. Nature Genetics, 44(8), 841-847.

[11] Wang, R. S., Marin, A., & Loscalzo, J. (2021). Network-based disease mechanisms and precision medicine applications. Frontiers in Genetics, 12, 679940

[12] Liu, R., Li, M., Wang, Y., Pan, Y., & Wang, S. (2020). Network-based approach to drug repositioning. Molecular BioSystems, 12(2), 314-323.

[13] Schubert, M., Klinger, B., Klünemann, M., et al. (2021). Causal networks for targeting cancer vulnerability. Briefings in Bioinformatics, 22(2), 179-191.

[14] Pearl, J. (1988). Probabilistic reasoning in intelligent systems: Networks of plausible inference. Morgan Kaufmann.

[15] Friedman, N., Linial, M., Nachman, I., & Pe'er, D. (2000). Using Bayesian networks to analyze expression data. Journal of Computational Biology, 7(3-4), 601-620.

[16] Kline, R. B. (2015). Principles and practice of structural equation modeling. Guilford publications.

[17]Shipley, B. (2000). A new inferential test for path models based on directed acyclic graphs. Structural Equation Modeling, 7(2), 206-218.

[18] Kauffman, S. A. (1993). The origins of order: Self-organization and selection in evolution. Oxford University Press.

[19] Albert, R., & Othmer, H. G. (2003). The topology of the regulatory interactions predicts the expression pattern of the segment polarity genes in Drosophila melanogaster. Journal of Theoretical Biology, 223(1), 1-18.

[20] Liu A., Trairatphisan P., Gjerga E. et al. From expression footprints to causal pathways: contextualizing large signaling networks with CARNIVAL npj Systems Biology and Applications volume 5, Article number: 40 (2019) (equal contributions).

[21] Melas IN, Sakellaropoulos T, Iorio F, Alexopoulos L, Loh WY, Lauffenburger DA, Saez-Rodriguez J, Bai JPF. (2015). Identification of drug-specific pathways based on gene expression data: application to drug induced lung injury. Integrative Biology, Issue 7, Pages 904-920,

[22] Schubert M, Klinger B, Klünemann M, Sieber A, Uhlitz F, Sauer S, Garnett MJ, Blüthgen N, Saez-Rodriguez J. (2018). Perturbation-response genes reveal signaling footprints in cancer gene expression. Nature Communication, Issue 9, Nr. 20.

[23].Pearl, J. (1985). Bayesian networks : A model of self-activated memory for evidential

Reasoning. University of California (Los Angeles). Computer Science Department.

[24] Madigan, D., York, J., and Allard, D. (1995). Bayesian graphical models for discrete

Data. International Statistical Review/Revue Internationale de Statistique, pages 215–

232. 24, 91

[25] Madigan, D. and York, J. C. (1997). Bayesian methods for estimation of the size of a

Closed population. Biometrika, 84(1) :19–31. 91

[26] A. Liu, P. Trairatphisan, E. Gjerga, et al., “From expression footprints to causal pathways: contextualizing large signaling networks with CARNIVAL,” npj Systems Biology and Applications, vol. 5, no. 40, 2019.

[27] “Analyse d’inférence causale (Statistiques spatiales)—ArcGIS Pro,” [En ligne]. Disponible sur : https://pro.arcgis.com/fr/pro-app/3.3/tool-reference/spatial-statistics/causal-inference-analysis.htm

[28] “Fonctionnement de l’analyse d’inférence causale—ArcGIS Pro,” [En ligne]. Disponible sur : https://pro.arcgis.com/fr/pro-app/3.3/tool-reference/spatial-statistics/how-causal-inference-analysis-works.htm

[29] J. Zhang et al., “High-throughput identification of cancer biomarkers in human body fluids,” Nature Communications, vol. 11, no. 1, p. 1, 2020.

[30] B. Schölkopf et al., “Toward causal representation learning,” Proceedings of the National Academy of Sciences, vol. 118, no. 48, p. e2018103118, 2021.

[31] D’haeseleer, P., Liang, S., and Somogyi, R. (2000). Genetic network inference : from

Co-expression clustering to reverse engineering. Bioinformatics, 16(8) :707–726. 10

[32] Emmert-Streib, F., Dehmer, M., and Haibe-Kains, B. (2014). Gene regulatory net-

Works and their applications : understanding biological and medical problems in terms

Of networks. Frontiers in cell and developmental biology, 2 :38. 9, 50

[33] Emmert-Streib, F., Glazko, G., De Matos Simoes, R., et al. (2012). Statistical inference

And reverse engineering of gene regulatory networks from observational expression

Data. Frontiers in genetics, 3 :8. 43

[34] Folmer, E. O., van der Geest, M., Jansen, E., Olff, H., Anderson, T. M., Piersma, T.,

And van Gils, J. A. (2012). Seagrass–sediment feedback : an exploration using a non-

Recursive structural equation model. Ecosystems, 15(8) :1380–1393. 30

[35] Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estima-

Tion with the graphical lasso. Biostatistics, 9(3) :432–441. 10, 17, 44, 58

[36] Friedman, N. and Koller, D. (2003). Being bayesian about network structure. A baye-

Sian approach to structure discovery in bayesian networks. Machine learning, 50(1-2) :95–125. 26, 76, 86

[37] Friedman, N., Linial, M., Nachman, I., and Pe’er, D. (2000). Using bayesian networks

To analyze expression data. Journal of computational biology, 7(3-4) :601–620. 77

[38]Shipley, B. (2000). A new inferential test for path models based on directed acyclic graphs. Structural Equation Modeling, 7(2), 206-218.

[39] Kauffman, S. A. (1993). The origins of order: Self-organization and selection in evolution. Oxford University Press.

[40] Albert, R., & Othmer, H. G. (2003). The topology of the regulatory interactions predicts the expression pattern of the segment polarity genes in Drosophila melanogaster. Journal of

Theoretical Biology, 223(1), 1-18.
[41] Illustration générée par IA pour représenter le processus d'inférence de réseaux causaux avec CARNIVAL, 2025.

[42] saezlab, "CARNIVAL: Causal Network Inference," GitHub Repository, [En ligne]. Disponible sur : https://github.com/saezlab/CARNIVAL. [Consulté le: avril 2025].

[42] Liu A, Trairatphisan P, Gjerga E, Didangelos A, Barratt J, Saez-Rodriguez J. From expression footprints to causal pathways: contextualizing large signaling networks with CARNIVAL. NPJ Systems Biology and Applications, 2019. [GitHub: https://github.com/saezlab/CARNIVAL]

[43] Türei D, Korcsmáros T, Saez-Rodriguez J. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. Nature Methods, 2016.

[44] He E, Wierling C, Lehrach H, Saez-Rodriguez J. Mathematical models of cancer signaling pathways. In Systems Medicine, Academic Press, 2016.

[45] Pearl, J. (2009). Causality: Models, reasoning, and inference. Cambridge university press.

[46] Margolin, A.A. et al. (2006). ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics.

[47] Huynh-Thu, V.A. et al. (2010). Inferring regulatory networks from expression data using tree-based methods. PLoS One.

[48] Müssel, C. et al. (2010). BoolNet—an R package for generation, reconstruction and analysis of Boolean networks. Bioinformatics.

[49] Holland, C.H. et al. (2020). Robustness and applicability of transcription factor and pathway analysis tools on single-cell RNA-seq data. Genome Biology.

[50] Türei, D. et al. (2016). OmniPath: guidelines and gateway for literature-curated signaling pathway resources. Nature Methods.

[51] Shannon, P. et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Research.

Annexes

Annexe 1 : 1. Prétraitement et Normalisation (R)

```
# تأكد من عدد الأعمدة في count_data
ncol(count_data)
# لازم يكون 100 هنا (مثلاً)

# توليد نوع العينة لكل عمود (50 Normal + 50 Tumor)
sample_type <- c(rep("Tumor", 50), rep("Normal", 50))

# إعداد col_data بنفس الترتيب
,col_data <- data.frame(row.names = colnames(count_data)
(sample_type = sample_type

# تحويل sample_type إلى عامل
col_data$sample_type <- as.factor(col_data$sample_type)

# إنشاء DESeqDataSet
,dds <- DESeqDataSetFromMatrix(countData = count_data
,colData = col_data
(design = ~ sample_type

dds <- DESeq(dds)
vsd <- vst(dds)
# إنشاء مجلد للصور
} if (!dir.exists("figures"))
dir.create("figures")
{
```

```
قبل و بعد التطبيع Boxplot #
npj(figures/fig4.1_boxplot.png", width=1200, height=600")
par(mfrow=c(1,2))
 ,boxplot(log2(count_data + 1)
 ,las=2
 ,"main="Avant Normalisation
 ,"col="lightgray
(cex.axis=0.7
 ,boxplot(assay(vsd)
 ,las=2
 ,"main="Après Normalisation (VST)
 ,"col="lightblue
(cex.axis=0.7
()dev.off
```

Annexe 2: Matrice d'activité des TFs

```
# tf_activity من الناتج هو VIPER

للمراجعة القيم وبعض الأبعاد طباعة #

dim(tf_activity)

head(tf_activity[, 1:5])  كمثال عينات 5 أول نشاط #
```

Annexe 3: Corrélation avec l'expression des gènes cibles

```
library(stats)

معين TF نختار #

tf_name <- rownames(tf_activity)[1]

TF لهذا الهدف جينات #

target_genes <- names(regulon_clean[[tf_name]]$tfmode)

التعبير بيانات في موجودة الجينات أن التأكد #
```

```
target_genes <- intersect(target_genes, rownames(expr_mat_filtered))

# نشاط TF عبر العينات

tf_scores <- tf_activity[tf_name, ]

# التعبير الجيني لجينات الهدف عبر العينات

expr_targets <- expr_mat_filtered[target_genes, , drop=FALSE]

# حساب متوسط التعبير للجينات الهدف عبر العينات

mean_expr_targets <- colMeans(expr_targets)

# حساب معامل الارتباط بيرسون بين نشاط TF وتعبير الهدف ومتوسط

correlation <- cor(tf_scores, mean_expr_targets, method = "pearson")

print(paste("Correlation for", tf_name, ":", correlation))
```

Annexe 4: pour le facteur de transcription MYC

```
correlations <- sapply(rownames(tf_activity), function(tf) {

  target_genes <- names(regulon_clean[[tf]]$tfmode)

  target_genes <- intersect(target_genes, rownames(expr_mat_filtered))

  if(length(target_genes) < 3) {

    return(NA)  # إذا جينات الهدف قليلة جدًا

  }

  tf_scores <- tf_activity[tf, ]

  expr_targets <- expr_mat_filtered[target_genes, , drop=FALSE]

  mean_expr_targets <- colMeans(expr_targets)



  cor(tf_scores, mean_expr_targets, method = "pearson")
```

```
})

# عرض عدد عوامل النسخ التي لها correlation > 0.7

sum(correlations > 0.7, na.rm = TRUE)
```

Annexe 5: Comparaison aux signatures connues dans MSigDB

```
library(msigdbr)

library(fgsea)  # تحليل التخصيص الغني

# تحميل التواقيع gene sets (مثلاً Hallmark gene sets)

msigdb_sets <- msigdbr(species = "Homo sapiens", category = "H")

# نجهز gene sets في شكل قائمة

gene_sets <- split(msigdb_sets$gene_symbol, msigdb_sets$gs_name)

# اختبار الارتباط أو تحليل GSEA بين tf_activity وأنشطة المجموعات

# مثال: تحليل إحصائي باستخدام fgsea أو correlation مباشر

# هنا ممكن تحاولي تحليلات تخصيص غنى (GSEA) لكل TF ضد نشاط gene sets
```

Annexe 6: Heatmap des scores d’activité des 30 TFs les plus actifs

```
# عدد اختصار يمكن TFs نشاطًا الأكثر (مثلاً أعلى 30)

top_tfs <- names(head(ranks, 30))

# استخراج بيانات النشاط لهذه TFs فقط

tf_activity_top <- tf_activity[top_tfs, ]

# رسم heatmap

pheatmap(tf_activity_top,

         scale = "row",

         clustering_distance_rows = "correlation",
```

```
    clustering_distance_cols = "correlation",
    main = "Heatmap of Top 30 TF Activity Scores")
```

Annexe7 : premières lignes du réseau d'interactions orientées et signées importé depuis OmniPath (consensus_direction = 1)

```
library(igraph)
# عندنا الرسم g
# حساب درجات العقد (Degree)
deg_in <- degree(g, mode = "in")
deg_out <- degree(g, mode = "out")
deg_total <- degree(g, mode = "all")
# إضافة درجات العقد كخصائص للعقد (nodes)
V(g)$deg_in <- deg_in
V(g)$deg_out <- deg_out
V(g)$deg_total <- deg_total
# حساب مركزية بين الوسيطية (Betweenness)
V(g)$betweenness <- betweenness(g)
# الكشف عن المجتمعات (clusters)
clusters <- cluster_walktrap(g)
V(g)$community <- membership(clusters)
# نظرة سريعة على أهم 5 جينات حسب الدرجة الكلية
head(sort(deg_total, decreasing = TRUE), 5)
```

Annexe8 : Visualisation du réseau causal inféré avec igraph

```
library(igraph)

edges <- results$weightedSIF

colnames(edges) <- c("source", "sign", "target", "weight")

g <- graph_from_data_frame(d = edges[, c("source", "target")], directed = TRUE)

E(g)$sign <- edges$sign

# الألوان حسب التحفيز (1 أزرق) أو تثبيط (-1 أحمر)

E(g)$color <- ifelse(E(g)$sign == 1, "blue", "red")

# ما يتجاهلش layout خليك تستخدم وزن ثابت عشان

E(g)$weight <- 1

# اختار layout مناسب مثل layout_with_fr (Force-directed)

plot(g, edge.color = E(g)$color, vertex.label.cex = 0.8, vertex.size = 15,

     layout = layout_with_fr(g))
```

Annexe 9 : Analyse topologique du réseau causal inféré à l'aide du package igraph, montrant le calcul des degrés d'entrée, de sortie et totaux, la centralité d'intermédiarité (betweenness), ainsi que la détection des communautés

```
library(igraph)

# عندنا الرسم g

# حساب درجات العقد (Degree)

deg_in <- degree(g, mode = "in")

deg_out <- degree(g, mode = "out")
```

```
deg_total <- degree(g, mode = "all")

# إضافة درجات العقد كخصائص للعقد (nodes)

V(g)$deg_in <- deg_in

V(g)$deg_out <- deg_out

V(g)$deg_total <- deg_total

# حساب مركزية بين الوسيطية (Betweenness)

V(g)$betweenness <- betweenness(g)

# الكشف عن المجتمعات (clusters)

clusters <- cluster_walktrap(g)

V(g)$community <- membership(clusters)

# نظرة سريعة على أهم 5 جينات حسب الدرجة الكلية

head(sort(deg_total, decreasing = TRUE), 5)
```

Annexe 10 : Réseau causal annoté avec les activités géniques (AvgAct), les relations causales et les mesures topologiques.

```
# أولاً، نجمع معلومات النشاط من nodesAttributes

nodes_attr <- results$nodesAttributes

# نجهز نشاط كل جين (نستخدم AvgAct)

activity <- setNames(nodes_attr$AvgAct, nodes_attr$Node)

# أضف نشاط للعقد في الرسم

V(g)$activity <- activity[V(g)$name]

# لون العقد حسب النشاط:
```

```
# up  أخضر، = (موجب)down رمادي = صفر، أحمر = (سالب)

V(g)$color <- ifelse(V(g)$activity > 0, "green",

                     ifelse(V(g)$activity < 0, "red", "gray"))



# (مثلاً بين 5 و 20) حجم العقد حسب الدرجة الكلية

V(g)$size <- scales::rescale(V(g)$deg_total, to = c(5, 20))

# لون الحواف (تحفيز/تثبيط) كما قبل

E(g)$color <- ifelse(E(g)$sign == 1, "blue", "red")

# رسم الشبكة مع تحسينات

plot(g,

    vertex.label.cex = 0.7,

    vertex.label.color = "black",

    edge.arrow.size = 0.4,

    edge.color = E(g)$color,

    vertex.color = V(g)$color,

    vertex.size = V(g)$size,

    layout = layout_with_kk)
```

Annexe11 : Les 10 gènes les plus connectés dans le réseau causal basé sur le degré total.

```
# لكل عقدة out-degree و in-degree حساب

deg_in <- degree(g, mode = "in")
```

```
deg_out <- degree(g, mode = "out")



# نجمعهم في جدول مع أسماء العقد

degree_df <- data.frame(

  node = names(deg_in),

  in_degree = deg_in,

  out_degree = deg_out,

  total_degree = deg_in + deg_out

)

# عرض أعلى 10 عقد حسب درجة التأثير (total degree)

head(degree_df[order(-degree_df$total_degree), ], 10)
```

Annexe 12 : Visualisation du réseau avec mise en évidence des gènes influents selon le degré total

```
# ربط درجات العقد بالرسم البياني

V(g)$degree <- degree_df$total_degree[match(V(g)$name, degree_df$node)]

# تحديد حجم العقد حسب درجة التأثير (مع مقياس)

V(g)$size <- scales::rescale(V(g)$degree, to = c(10, 30))  # الحجم بين 10 و 30

# تحديد لون العقد (مثلاً بدرجات الأزرق حسب درجة التأثير)

library(RColorBrewer)

colors <- colorRampPalette(brewer.pal(9, "Blues"))(length(V(g)))

V(g)$color <- colors[rank(-V(g)$degree)]
```

```
# رسم الشبكة مع الأحجام والألوان الجديدة

plot(g,

    edge.color = E(g)$color,

    vertex.label.cex = 0.7,

    vertex.label.color = "black",

    vertex.frame.color = "gray",

    layout = layout_with_kk(g)

)
```

Annexe 13 : Représentation du réseau causal avec codage couleur basé sur l'activité des gènes.

```
# نأخذ جدول النشاط

node_data <- results$nodesAttributes

# نحط أسماء العقد كـ row names باش نسهّل الوصول

rownames(node_data) <- node_data$Node

# دالة بسيطة لتحديد اللون حسب النشاط

get_node_color <- function(up, down) {

  if (up > 50) return("green")

  else if (down > 50) return("red")

  else return("lightgray")

}

# نطبّقها على كل عقدة في الرسم البياني
```

```
V(g)$color <- sapply(V(g)$name, function(n) {
  if (n %in% node_data$Node) {
    get_node_color(node_data[n, "UpAct"], node_data[n, "DownAct"])
  } else {
    "lightgray"  # لعقد غير موجودة في node_data
  }
})
plot(g,
     edge.color = E(g)$color,
     vertex.color = V(g)$color,
     vertex.label.cex = 0.7,
     vertex.label.color = "black",
     vertex.frame.color = "gray",
     layout = layout_with_fr(g),
     main = "شبكة CARNIVAL بألوان النشاط الجيني"
)
```

Annexe 14 : Visualisation du réseau causal coloré par communauté (méthode de Louvain).

```
# ضروري يكون الرسم غير موجه، نحولو مؤقتًا
g_undirected <- as.undirected(g, mode = "collapse")
# نستخدم خوارزمية Louvain
```

```
communities <- cluster_louvain(g_undirected)

# عدد المجموعات

length(communities)

# العقد في كل مجموعة

sizes(communities)

# عرض أول 5 مجموعات وعقدها

for (i in 1:5) {

  cat(paste("Community", i, ":", "\n"))

  print(communities[[i]])

  cat("\n")

}

# عدد الألوان المطلوبة

n <- length(communities)

# توليد ألوان مميزة

colors <- rainbow(n)

# نلون العقد حسب المجموعة

V(g)$color <- colors[membership(communities)[V(g)$name]]

# رسم الشبكة مع المجتمعات

plot(g,

     edge.color = E(g)$color,

     vertex.color = V(g)$color,

     vertex.label.cex = 0.7,
```

```
    vertex.label.color = "black",
    vertex.frame.color = "gray",
    layout = layout_with_fr(g),
    main = "شبكة CARNIVAL ملوّنة حسب المجتمعات (Communities)"
)
# PNG
png("network_communities.png", width = 1200, height = 1000)
plot(g,
    edge.color = E(g)$color,
    vertex.color = V(g)$color,
    vertex.label.cex = 0.7,
    vertex.label.color = "black",
    vertex.frame.color = "gray",
    layout = layout_with_fr(g),
    main = "شبكة CARNIVAL ملوّنة حسب المجتمعات"
)
dev.off()
# PDF
pdf("network_communities.pdf", width = 10, height = 8)
plot(g,
    edge.color = E(g)$color,
    vertex.color = V(g)$color,
```

```
    vertex.label.cex = 0.7,
    vertex.label.color = "black",
    vertex.frame.color = "gray",
    layout = layout_with_fr(g),
    main = "ملوّنة حسب المجتمعات CARNIVAL شبكة"
)
dev.off()
```

Annexe 15 : Liste des gènes et des communautés associées dans le réseau d'interactions

```
# استخراج عضوية كل عقدة (gene → رقم المجتمع)
membership_df <- data.frame(
  gene = names(membership(communities)),
  community = membership(communities)
# نعاين أول السطور
head(membership_df)
write.csv(membership_df, "gene_communities.csv", row.names = FALSE)
```

Annexe 16 : l'activité moyenne par communauté

```
# نربط nodesAttributes مع المجتمع
merged_data <- merge(results$nodesAttributes, membership_df, by.x = "Node", by.y = "gene")
# نحسب المتوسط لكل مجتمع
```

```
avg_activity_by_comm <- aggregate(AvgAct ~ community, data = merged_data, FUN = mean)

# نعرضهم

print(avg_activity_by_comm)
```

Annexe 17 : Répartition du nombre de gènes par communauté dans le réseau CARNIVAL

```
library(ggplot2)

# عدد الجينات في كل مجتمع

gene_counts <- as.data.frame(table(membership_df$community))

colnames(gene_counts) <- c("Community", "GeneCount")

# رسم بياني

ggplot(gene_counts, aes(x = factor(Community), y = GeneCount)) +

  geom_bar(stat = "identity", fill = "skyblue") +

  labs(title = "عدد الجينات في كل مجتمع", x = "المجتمع", y = "عدد الجينات") +

  theme_minimal()

pdf("gene_communities_barplot.pdf", width = 8, height = 6)

ggplot(gene_counts, aes(x = factor(Community), y = GeneCount)) +

  geom_bar(stat = "identity", fill = "skyblue") +

  labs(title = "عدد الجينات في كل مجتمع", x = "المجتمع", y = "عدد الجينات") +

  theme_minimal()

dev.off()
```

Annexe 18 : Sous-graphe du groupe de gènes le plus actif identifié par CARNIVAL

```
# نرتب المجتمعات حسب النشاط المتوسط تنازليًا

most_active_comm <- avg_activity_by_comm[order(-avg_activity_by_comm$AvgAct), ]

# نعرض المجتمع الأكثر نشاطًا

head(most_active_comm, 1)

# ناخذ رقم المجتمع الأكثر نشاطًا

top_comm_id <- most_active_comm$community[1]

# نختار الجينات المنتمية لهذا المجتمع

top_comm_genes <- merged_data[merged_data$community == top_comm_id, ]



# نعرض أهم الجينات في هذا المجتمع حسب النشاط

top_genes_sorted <- top_comm_genes[order(-top_comm_genes$AvgAct), ]

# نعرض أول 10

head(top_genes_sorted, 10)[, c("Node", "AvgAct", "UpAct", "DownAct")]

# حفظ الجينات داخل المجتمع الأكثر نشاطًا

write.csv(top_genes_sorted, "top_active_community_genes.csv", row.names = FALSE)



# حفظ معلومات المجتمع الأكثر نشاطًا

write.csv(most_active_comm, "communities_avg_activity_sorted.csv", row.names = FALSE)

library(igraph)

# استخراج أسماء الجينات في المجتمع الأكثر نشاطًا
```

```
top_nodes <- top_genes_sorted$Node

# استخراج subgraph يحتوي فقط على هاد الجينات

sub_g <- induced_subgraph(g, vids = V(g)[name %in% top_nodes])

# رسم الشبكة المصغّرة

plot(

  sub_g,

  main = "أكثر مجتمع نشاطًا في الشبكة",

  vertex.label.cex = 0.8,

  vertex.size = 20,

  edge.color = E(sub_g)$color,

  vertex.label.color = "black",

  layout = layout_with_fr

)

# نحفظ الصورة PNG

png("subgraph_top_active_community.png", width = 1000, height = 800)

plot(

  sub_g,

  main = "أكثر مجتمع نشاطًا في الشبكة",

  vertex.label.cex = 0.8,

  vertex.size = 20,

  edge.color = E(sub_g)$color,

  vertex.label.color = "black",
```

```
  layout = layout_with_fr
)
dev.off()


# نحفظها أيضًا PDF
pdf("subgraph_top_active_community.pdf", width = 10, height = 8)
plot(
  sub_g,
  main = "أكثر مجتمع نشاطًا في الشبكة",
  vertex.label.cex = 0.8,
  vertex.size = 20,
  edge.color = E(sub_g)$color,
  vertex.label.color = "black",
  layout = layout_with_fr
)
dev.off()
```

Résumé

Ce mémoire présente une étude approfondie de l'analyse des réseaux causaux dans le cancer à l'aide de la méthodologie CARNIVAL, en utilisant des techniques informatiques avancées pour décrypter la complexité des réseaux moléculaires et identifier les nœuds régulateurs clés. Grâce à l'analyse des données transcriptomiques et à l'application d'algorithmes de détection de communautés, nous avons identifié des gènes centraux et des communautés géniques hautement actives pouvant représenter des vulnérabilités thérapeutiques prometteuses. Malgré les défis liés à la qualité des données et à la nécessité d'une validation expérimentale, cette recherche ouvre des perspectives importantes pour comprendre les mécanismes du cancer et développer des thérapies ciblées, tout en soulignant le potentiel d'intégration avec des données multi-omiques et des applications de médecine personnalisée à l'avenir.

Abstract

This thesis provides a comprehensive study of causal network analysis in cancer using the CARNIVAL methodology, employing advanced computational techniques to decipher the complexity of molecular networks and identify key regulatory nodes. Through transcriptomic data analysis and community detection algorithms, we identified central genes and highly active gene communities that may represent promising therapeutic vulnerabilities. Despite challenges related to data quality and the need for experimental validation, this research opens important avenues for understanding cancer mechanisms and developing targeted therapies, while highlighting the potential for integration with multi-omics data and personalized medicine applications in the future.

ملخص

تقدم هذه المذكرة البحثية دراسة متكاملة لتحليل الشبكات السببية في السرطان باستخدام منهجيةCARNIVAL ، حيث توظف تقنيات حسابية متقدمة لفك تعقيدات الشبكات الجزيئية وتحديد العقد التنظيمية الرئيسية. من خلال تحليل البيانات النسخية وتطبيق خوارزميات اكتشاف المجتمعات، تمكنا من تحديد جينات مركزية ومجتمعات جينية عالية النشاط قد تشكل نقاط ضعف علاجية واعدة. ورغم التحديات المتعلقة بجودة البيانات وضرورة التحقق التجريبي، تفتح هذه الدراسة آفاقاً مهمة في فهم آليات السرطان وتطوير علاجات مستهدفة، كما تبرز إمكانات كبيرة للتكامل مع البيانات متعددة الوسوم وتطبيقات الطب الشخصي في المستقبل.