UNIVERSITÉ NATIONALE DU VIETNAM, HANOI

Détermination du rôle de la protéine SigmaR1 dans le cancer du pancréas par une approche d'intelligence artificielle

MÉMOIRE DE FIN D'ÉTUDES DE MASTER EN INFORMATIQUE

Spécialité : Systèmes Intelligents et Multimédias

Redigé par : Sous la direction de :

2023

Contents

1 Introduction

Les dernières décennies ont été marquées par une évolution significative dans la manière dont les sciences sont abordées. Alors qu'auparavant, les disciplines scientifiques étaient souvent traitées de manière indépendante, la tendance actuelle est à la transdisciplinarité. Un exemple de cette convergence entre différentes disciplines est la bioinformatique, qui réunit la biologie, l'informatique et les statistiques mathématiques.

1.1 Contexte

La bioinformatique représente un mariage harmonieux entre la biologie et l'informatique, exploitant les outils statistiques et la modélisation informatique au service de la recherche biologique. Elle se définit par le développement et l'application de méthodes informatiques et statistiques visant à comprendre et interpréter les informations génétiques et moléculaires. Cette approche transcende les frontières traditionnelles des disciplines, offrant ainsi un moyen puissant d'aborder les complexités des données biologiques à grande échelle.

Au cœur de la bioinformatique se trouve son rôle essentiel dans l'analyse des données génomiques. Elle va au-delà de la simple manipulation de séquences génétiques pour devenir un moteur de création de cadres expérimentaux pour les biologistes. En fournissant des outils, des méthodes et des ressources indispensables, la bioinformatique facilite la planification, l'exécution et l'analyse d'expériences biologiques.

Dans ces cadres expérimentaux, les biologistes s'emploient à mesurer l'activité des gènes pour identifier des gènes ou groupes de gènes activés, potentiellement liés aux phénotypes observés. Ces analyses, basées sur la mesure de l'activité génique, visent à comprendre les relations entre les caractéristiques observées (phénotypes) et à identifier des schémas susceptibles de devenir des cibles thérapeutiques.

Notre projet, en collaboration étroite avec des biologistes, illustre cette convergence. L'objectif premier de cette collaboration est de déterminer le rôle de la protéine SigmaR1 dans le contexte spécifique du cancer du pancréas. Bien que le rôle précis de cette protéine reste à élucider, son étude revêt un intérêt particulier dans la compréhension des mécanismes génétiques liés à cette forme de cancer pour le développement de thérapies plus ciblées et efficaces.

1.2 Problématique

L'évolution vers la transdisciplinarité dans le domaine scientifique, incarnée par la bioinformatique, suscite des questionnements essentiels sur la manière dont nous abordons la recherche et la compréhension des phénomènes biologiques. Face à la complexité croissante des données génomiques, la bioinformatique doit relever plusieurs défis, appelant à une réévaluation des méthodes et des approches utilisées.

La complexité inhérente aux données génomiques à grande échelle constitue l'un des principaux défis pour la bioinformatique. Malgré ses outils avancés pour manipuler et interpréter ces données, la sélection des gènes à étudier par les biologistes demeure un enjeu crucial. La méthode conventionnelle consiste à rechercher des modules de gènes dont l'action combinée réalise une fonction spécifique, et cela se fait souvent à travers l'approche des **'top k-gènes'** les plus variables [Rap+07]. Cependant, cette méthode montre ses limites, car la variabilité ne garantit pas toujours la pertinence biologique. Cette limite peut parfois être observée dans le cas des gènes inflammatoires, où les gènes les plus variables représentent souvent des causes plutôt que des facteurs directement liés aux observations. Ce défi devient particulièrement pressant lorsque l'on cherche à démêler les mécanismes d'une fonction biologique Comment facilement identifier ces modules de gènes qui varient dans l'expérience mais qui interagissent également entre eux pour une fonction biologique associée ?

La recherche de modules de gènes actifs, similaire à la détection de communautés dans les réseaux sociaux, s'avère cruciale pour comprendre les interactions génétiques. Toutefois, contrairement aux réseaux sociaux où les attributs des individus raffinent les communautés, en bioinformatique, la valeur associée aux gènes est aussi cruciale que la topologie du réseau. La recherche de modules nécessite des méthodes innovantes intégrant à la fois la topologie du graphe et la valeur des nœuds. Comment parvenir à une intégration efficace de ces deux aspects pour dévoiler des modules d'activité génique pertinents ? ou Comment réussir à fusionner de manière efficace la topologie des réseaux génétiques avec la valeur associée aux gènes pour dévoiler des modules actifs significatifs?

L'embedding de réseau, à travers une représentation vectorielle dense en dimension réduite, émerge comme une solution prometteuse dans cette quête. Un autre aspect crucial réside dans la dynamique des valeurs des nœuds dans les réseaux biologiques. Alors que la topologie des réseaux reste stable, les valeurs des nœuds peuvent évoluer à chaque nouvelle expérience. Comment adapter les méthodes d'embedding de réseau pour garantir une représentation fidèle aux données, prenant en compte cette dynamique propre à la biologie expérimentale et à la variabilité des valeurs des nœuds?

Notre parcours de recherche s'anime autour de ces questionnements complexes, avec pour objectif d'explorer de nouvelles perspectives dans l'utilisation de l'embedding multivues en bioinformatique. En répondant à ces défis, nous espérons contribuer à une compréhension approfondie de la recherche des modules actifs des reseaux génétiques, ouvrant la voie à des avancées significatives dans le domaine des thérapies ciblées.

1.3 Motivation du projet

Notre quête pour une compréhension approfondie des mécanismes génétiques est motivée par la nécessité de développer une méthode générique pour la détection de modules actifs. Dans le cadre de ce projet, cette méthode sera appliquée au contexte complexe du cancer du pancréas, une maladie dévastatrice souvent détectée en phase avancée mais elle pourra être également utilisée pour traiter de nombreux autres jeux de données produits par les biologistes. Les méthodes conventionnelles pour identifier les modèles de gènes qui varient dans l'expérience mais qui interagissent également entre eux pour une fonction peuvent parfois se révéler insuffisantes face à la complexité des activités biologiques.

Notre aspiration est de dépasser les limites des méthodes conventionnelles en explorant des approches plus adaptées pour faire des choix initiaux plus judicieux. Nous aspirons à contribuer à la recherche de méthodologies avancées permettant de discerner avec précision les gènes actifs impliqués dans des activités biologiques complexes. Notre projet vise à devenir un élément clé dans la recherche en bioinformatique, en utilisant l'embedding multivue. Cette technique, qui exploite la flexibilité de l'utilisation de plusieurs vues de données, promet une meilleure compréhension des mécanismes génétiques.

Dans notre projet, nous adoptons une approche multi-vue où un réseau d'interaction est considéré comme une vue et un ensemble de mesures effectuées comme une autre. Cette stratégie permet l'intégration de multiples perspectives de données, offrant une flexibilité accrue. En cas de succès avec deux vues - un graphe d'interaction et des valeurs d'expression - nous envisageons d'étendre la méthode en intégrant d'autres réseaux ou mesures. Cette stratégie est d'autant plus pertinente dans le contexte biologique où la topologie du réseau reste stable, tandis que les valeurs des nœuds varient avec chaque nouvelle expérience.

Notre intérêt est de réaliser un embedding multivues, où chaque série d'expériences pourrait être représentée par des graphes basés sur les valeurs des nœuds, formant ainsi plusieurs réseaux. Cette approche pourrait améliorer significativement notre compréhension et notre analyse des données dans des environnements biologiques complexes, ouvrant la voie à de nouvelles perspectives dans l'étude du multivue.

En somme, notre motivation repose sur la volonté de transcender les limitations actuelles et d'explorer des voies novatrices en bioinformatique. En utilisant des techniques avancées comme l'embedding multivue et le deep learning, nous visons à repousser les frontières de la recherche, en allant au-delà de la simple sélection des "top k-gènes" par leur variation. Notre but ultime est de mieux comprendre les mécanismes sous-jacents complexes entre les gènes et ainsi orienter la recherche vers des cibles thérapeutiques plus pertinentes.

1.4 Objectifs de la recherche

Notre recherche a pour objectif principal d'explorer et d'évaluer l'efficacité de la détection de modules actifs en utilisant des données multivue, en exploitant les techniques d'apprentissage profond offertes par le deep

learning pour former des embeddings en dimension réduite. Cette approche novatrice vise à surmonter les limitations de l'approche traditionnelle, qui combine la structure du graphe génétique avec la valeur des nœuds, à l'instar de la méthode AMINE[Pas+22]. Nous envisageons également de comparer ces résultats à ceux obtenus par la méthode AMINE, reconnue pour ses performances supérieures dans la détection de modules actifs.

La flexibilité de l'approche multivue constitue le pivot central de notre investigation. Comme souligné dans notre motivation, cette méthode ouvre la porte à l'utilisation de plusieurs perspectives de données, telles que différents réseaux génétiques ou les résultats d'expériences réalisées à différents moments. En exploitant cette diversité de points de vue, la séparation des vues pourrait offrir des avantages significatifs en termes d'efficacité de calcul. Avec cette approche, nous avons l'opportunité de combiner plusieurs vues représentant des expériences distinctes. En cas de résultats supérieurs à ceux d'AMINE, notre deuxième objectif consistera à démontrer que cette approche permettra une représentation plus précise et adaptable des mécanismes génétiques, offrant ainsi une perspective plus complète pour la détection de modules actifs, compte tenu de l'interdépendance des activités biologiques.

En résumé, nos objectifs de recherche sont les suivants :

- Explorer l'efficacité de la détection de modules actifs en utilisant des données multivue: Nous souhaitons évaluer la capacité de l'embedding multivue à identifier de manière précise et complète les modules actifs dans les réseaux génétiques.

- Comparer les résultats avec la méthode AMINE : En confrontant les performances de notre approche à celles d'AMINE, reconnue pour ses succès dans la détection de modules actifs, nous cherchons à évaluer le potentiel de notre méthode à surpasser les approches existantes.

- Démontrer l'efficacité et l'adaptabilité de l'approche multivue : Si nos résultats confirment la supériorité de l'embedding multivue, notre objectif sera de démontrer comment cette approche offre une représentation plus précise et adaptable des mécanismes génétiques, fournissant ainsi une perspective plus complète pour la détection de modules actifs dans des conditions biologiques variées.

En poursuivant ces objectifs, notre ambition est de contribuer significativement à la recherche en bioinformatique et d'ouvrir de nouvelles voies pour une compréhension approfondie des mécanismes génétiques, particulièrement dans le contexte complexe du cancer du pancréas.

1.5 Questions de recherche

Notre parcours de recherche est guidé par des questionnements complexes visant à explorer les possibilités novatrices de l'embedding multivue en bioinformatique, en vue de générer une méthode générique applicable au cas particulier du cancer du pancréas. Ces interrogations émergent des lacunes identifiées dans

la recherche actuelle et cherchent à éclairer les défis spécifiques liés à l'analyse des réseaux génétiques à grande échelle.

- Comment l'embedding multivue peut-il optimiser la détection de modules actifs dans les réseaux génétiques par rapport à l'approche combinant la topologie du graphe et la valeur des nœuds?
- Quels critères de construction adopter pour élaborer le graphe de la vue dépendant des valeurs de poids des nœuds, afin de mieux conserver l'information dans l'embedding multivue?
- Quels critères de collaboration entre les vues devrions-nous mettre en exergue pour implémenter l'embedding multivue de manière optimale?
- Quels sont les avantages et les limitations de l'embedding multivue par rapport à la méthode AMINE, largement reconnue pour ses performances dans la détection de modules actifs?

En répondant à ces questions de recherche, notre ambition est d'apporter des contributions significatives à la compréhension des mécanismes génétiques, de développer des approches innovantes en bioinformatique, et de jeter les bases pour des avancées substantielles dans le domaine des thérapies ciblées. Notre objectif ultime est de mettre en place une méthode générique pour la détection des modules actifs dans les réseaux génétiques, que nous appliquerons particulièrement dans le contexte complexe du cancer du pancréas. En combinant l'embedding multivue avec des critères de construction et de collaboration judicieux, nous espérons ouvrir de nouvelles perspectives pour une meilleure compréhension des processus biologiques et contribuer ainsi à l'élaboration de thérapies plus précises et efficaces.

2

Chapter 2: Etat de l'Art

2.1 Introduction

Dans l'analyse et le traitement des données, il est crucial de comprendre les interactions et les corrélations entre différents ensembles de données. L'approche des embeddings multi-vues se distingue par son efficacité et sa flexibilité, jouant un rôle essentiel dans l'intégration d'ensembles de données hétérogènes. Cette méthode offre une représentation unifiée et globale qui saisit avec précision la complexité des données. Cette capacité est particulièrement utile pour extraire des informations pertinentes lors du traitement de données, que ce soit pour le regroupement (clustering), la classification ou la prédiction de liens. Compte tenu de l'importance d'intégrer les données dans un espace unifié, de nombreuses théories ont été élaborées, donnant naissance à des modèles à la fois robustes et efficaces.

2.2 Théories et modèles existants

2.2.1 Notion de Multivues

La notion de multivues fait référence à une approche analytique dans laquelle plusieurs "vues" ou représentations d'un même ensemble de données sont examinées et intégrées pour obtenir une compréhension plus profonde et détaillée. Cette méthode est extrêmement utile dans les domaines caractérisés par des données complexes et multidimensionnelles. Prenons l'exemple de la bioinformatique : une vue peut représenter des données génomiques, tandis qu'une autre pourrait se concentrer sur les données protéomiques ou transcriptomiques du même échantillon biologique. L'approche multivues permet d'analyser ces différentes perspectives simultanément, offrant ainsi une compréhension plus complète et nuancée du sujet étudié. Cette méthode diffère fondamentalement de l'approche univue, qui se limite à une seule perspective des données. En intégrant diverses vues, les chercheurs peuvent identifier des corrélations et des interactions qui ne seraient pas évidentes en examinant les vues séparément. Cela conduit à des insights plus riches et ouvre la voie à de nouvelles découvertes dans leur domaine de recherche.

L'importance d'intégrer diverses sources d'information dans un cadre unifié a conduit au développe-

ment de nombreuses théories, parfois basées sur des principes statistiques et probabilistes. Parmi cet éventail de concepts, quatre ont particulièrement retenu notre attention, chacun apportant une perspective unique et enrichissante à notre compréhension du multivues. Ces théories nous permettent de naviguer et d'exploiter efficacement la complexité inhérente aux données multidimensionnelles, ouvrant ainsi de nouvelles voies dans la recherche et l'analyse de données.

2.2.2 Factorisation Matricielle par des Composants Liés pour l'Intégration Unifié

Dans leur approche de la "Factorisation Matricielle par des Composants Liés pour l'Intégration Unifiée",[Iri17] apportent une contribution notable avec leur modèle SLIDE (Structural Learning and Integrative DEcomposition). Ce modèle se distingue par l'intégration de structures partiellement partagées dans la factorisation matricielle des données multivues, une avancée par rapport au modèle JIVE (Joint and Individual Variation Explained)[Loc+13]. En fait dans la pluspart des un ensembles multivues reels , il existe des donnees dont les instances ne sont pas presents dans toute les vues .

SLIDE offre une représentation efficace des données multivues à travers des composants liés, utilisée pour la réduction dimensionnelle exploratoire et l'analyse d'association entre les vues. Cette intégration de composants partiellement partagés aborde un défi important dans la factorisation structurale des données multivues. Dans les études empiriques, notamment avec des données sur le cancer issues du répertoire "The Cancer Genome Atlas", Le modele SLIDE a démontré d'excellentes performances en termes d'estimation du signal et de sélection des composants.

Cependant, SLIDE présente des limitations. La méthode de détermination du nombre de composants pour chaque type (partagés, individuels, partiellement partagés) pour chaque vue, bien qu'innovante, peut rencontrer des difficultés en termes de complexité computationnelle et de précision dans des contextes de données variés. De plus, l'utilisation d'un cadre de factorisation matricielle pénalisée pour réduire la complexité peut limiter la flexibilité et l'adaptabilité du modèle dans certaines applications.

Ces limitations ouvrent la voie à l'exploration de nouvelles théories pour unifier les données multivues dans un cadre unifié, suggérant la nécessité de modèles plus flexibles et adaptatifs.

2.2.3 La Théorie Basée sur les cadre de probabiliste

La transformation des données issues de différentes vues dans un espace unifié, appelé espace partagé, est un processus fondamental dans le domaine de l'apprentissage automatique. Le modèle **Probabilistic Multi-view Graph Embedding (PMvGE)**[OS18]offre une solution innovante à ce défi, en combinant des techniques avancées d'incorporation de graphes (graph embedding) avec des approches probabilistes pour unifier les données multivues dans un espace commun.

Dans le cadre de PMvGE, les données de chaque vue sont initialement transformées en vecteurs de caractéristiques au sein d'un espace partagé. Cette étape est essentielle pour aligner les données provenant

de sources diverses. Elle est effectuée à l'aide de réseaux neuronaux, où les données d'entrée $x^{(v)}$ sont converties en vecteurs de caractéristiques $y^{(v)}$ dans l'espace partagé selon la fonction

$$y^{(v)} = f^{(v)}(x^{(v)}; \theta^{(v)}) \tag{2.1}$$

Le cœur de PMvGE réside dans sa capacité à modéliser la probabilité d'association entre des paires de vecteurs de caractéristiques issus de différentes vues. Cette probabilité est estimée par le produit scalaire des vecteurs de caractéristiques, exprimé par

$$P(y^{(v)}, y^{(u)}) = \sigma(y^{(v)T} y^{(u)}) \tag{2.2}$$

Cette méthode probabiliste dans le cas de l'implementation pourrai etre une sigmoid , cela permettrai non seulement d'identifier les associations entre les données, mais aussi d'en quantifier la force.

L'approche probabiliste pour modéliser les associations entre les vecteurs de caractéristiques est un élément clé de PMvGE. Elle permet de détecter et de mesurer la force des associations entre les données de différentes vues, offrant ainsi une compréhension plus détaillée des relations entre les données. L'objectif principal de PMvGE est de maximiser la vraisemblance des associations observées dans les données multivues. La fonction de vraisemblance $L(\theta)$ est optimisée pour ajuster les paramètres du modèle, suivant l'équation :

$$L(\theta) = \sum_{v,u} \sum_{i,j} w_{ij}^{(vu)} \log P(y_i^{(v)}, y_j^{(u)}; \theta) + (1 - w_{ij}^{(vu)}) \log(1 - P(y_i^{(v)}, y_j^{(u)}; \theta)) \tag{2.3}$$

Cette optimisation est cruciale pour assurer que le modèle reflète fidèlement les relations complexes entre les données multivue pour ce distingue des autres mmodeles. **PMvGE** (Probabilistic Multi-view Graph Embedding) se distingue de CDMCA par l'introduction de transformations non linéaires. Les réseaux neuronaux employe par PMvGE transformer les données de chaque vue en vecteurs de caractéristiques dans un espace partagé. Cette approche non linéaire permet à PMvGE de capturer des associations plus complexes et subtiles entre les vues, surpassant ainsi les capacités de CDMCA. En exploitant des concepts d'incorporation de graphes et de modélisation probabiliste, PMvGE facilite une analyse approfondie et une compréhension des données multivues. Toutefois, la transformation des données via des réseaux neuronaux peut être complexe et nécessite un ajustement précis des paramètres. De plus, la performance du modèle peut être limitée par la taille et la qualité des données disponibles.

Ces limitations soulignent l'importance de développer de nouvelles méthodes pour l'intégration de données multivues, incluant des techniques d'apprentissage plus simples

2.2.4 La Théorie Basée sur les Méthodes de Conflation de distribution de probabilité:

Le concept de conflation de distributions de probabilité joue un rôle essentiel dans la création d'embeddings unifiés. Cette approche, axée sur l'intégration et la fusion de données multivues en une représentation

unifiée, repose sur deux principes fondamentaux : la symétrie probabiliste et la réduction de la divergence de Kullback-Leibler. Dans ce contexte, chaque vue possède sa propre distribution de probabilité. Ces distributions sont calculées avant la fusion des données. Après cette fusion, une nouvelle distribution est obtenue dans l'espace unifié. Cela contraste avec les modèles probabilistes qui calculent la probabilité directement dans l'espace unifié.

1. Probabilité Symétrique

    - Chaque point dans un ensemble de données, relatif à une vue spécifique, calcule une probabilité symétrique pour chaque autre point potentiellement voisin.
    - La probabilité, notée p_{ij}^{v}, est déterminée par une formule basée sur l'exponentielle de la dissimilarité au carré entre les points, divisée par la somme de ces exponentielles pour tous les voisins potentiels.
    - La formule spécifique est :

    $$p_{ij}^{v} = \frac{\exp(-d_{ij}^{v})^2}{\sum_k \exp(-d_{ik}^{v})^2} \tag{2.4}$$

    où d_{ij}^{v} désigne la dissimilarité entre les échantillons i et j dans la vue v.

2. Réduction de la Divergence de Kullback-Leibler

    - L'objectif est de minimiser la divergence de Kullback-Leibler entre la distribution de probabilité dans l'espace de haute dimension et sa contrepartie dans l'espace de basse dimension (embedding).
    - Cette minimisation est cruciale pour optimiser l'embedding afin qu'il représente fidèlement les relations de probabilité des points dans l'espace de haute dimension.
    - La divergence est exprimée par :

    $$C = KL(P||Q) = \sum_{i,j} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

    .

Ce processus utilise des principes de probabilité et d'optimisation pour transformer efficacement des données multi-vues de haute dimension en un espace unifié de basse dimension, préservant ainsi la structure essentielle des données originelles.

2.2.5 La Théorie Basée sur les Méthodes de Collaboration de données:

Cette approche se concentre sur l'analyse collaborative des données afin de construire un embedding unifié. Elle repose sur trois propriétés essentielles pour explorer les interactions et relations entre les données dans

différentes vues. Ces propriétés sont : la diversité, la collaboration de premier ordre et la collaboration de second ordre

1. **Diversité**

   Cette phase vise à capturer l'unicité de chaque vue en produisant des paires d'échantillons ou individus qui sont réellement connectés au sein de chaque vue. Ces paires illustrent la similarité entre les échantillons dans une vue donnée. Pour une vue v, un ensemble de paires intra-vue $\Omega(v)$ est constitué, chaque paire $(u(v), w(v)) \in \Omega(v)$ comprenant un échantillon central $u(v)$ et un échantillon contextuel $w(v)$. L'objectif est d'optimiser la probabilité de prédire l'échantillon contextuel à partir de l'échantillon central, en réduisant la perte $Div(\Theta)$, définie comme :

$$L_{Div}(\Theta) = -\sum_{v \in V} \sum_{(i(v), j(v)) \in \Omega(v)} \log P(j(v)|i(v);\Theta) \tag{2.5}$$

2. **Collaboration de Premier Ordre**

   Bien que les différentes vues d'un réseau multi-vues présentent de la diversité, elles convergent finalement vers un ensemble commun d'individu ou echantillon. Les instances d'un même individu à travers différentes vues décrivent fondamentalement la même entité. Cette collaboration de premier ordre vise à aligner les représentations spécifiques d'un même échantillon à travers différentes vues. Pour ce faire, des paires intra-échantillon sont formées pour toute les vues dans lequelle l'instance de l'echantillon existe. Cette relation peut etre vue comme une relation identitaire car chaque paire représente le même individu observé dans différentes vues. Ainsi, nous avons les paires $(u(v), u(v'))$, où

   - $u(v)$ est l'échantillon dans la vue v
   - $u(v')$ est l'échantillon dans la vue v'.

   Comme il s'agit du même individu observé sous différentes perspectives, la perte $C1(\Theta)$ est minimiser pour maximiser la similarité des représentations vectorielles de l'échantillon dans les différentes vues. Cette perte $C1(\Theta)$ est exprimée de la manière suivante :

$$L_{C1}(\Theta) = -\sum_{v \in V} \sum_{(i(v), \cdot) \in \Omega(v)} \sum_{v' \neq v} \log P(i(v')|i(v);\Theta) \tag{2.6}$$

   ou

   - $P(i(v')|i(v);\Theta)$ exprime la probabilité de prédire correctement la représentation d'un nœud dans une vue v', en fonction de sa représentation dans une vue v, sous les paramètres Θ.
   - $i(v)$ et $i(v')$ indiquent respectivement le nœud central dans la vue v et sa représentation dans une autre vue v'.

L'objectif de cette fonction de perte $L_{C1}(\Theta)$ est de garantir que les embeddings d'un même nœud soient similaires à travers les différentes vues. En optimisant cette fonction, le modèle aligne efficacement les représentations de chaque nœud à travers les vues, assurant ainsi que les caractéristiques fondamentales du nœud sont cohérentes et fidèlement représentées dans l'ensemble du réseau multi-vues.

3. **Collaboration de Second Ordre**

   Cette collaboration utilise les associations entre les échantillons d'une vue pour améliorer la collaboration entre différentes vues. Des paires d'échantillons croisées sont établies selon les associations entre les échantillons de chaque vue, dans le but de mettre à jour les représentations d'un échantillon pour qu'elles ressemblent à celles des échantillons associés dans une autre vue. La perte $C2(\Theta)$ est minimisée et formulée comme suit :

$$L_{C2}(\Theta) = -\sum_{v\in V}\sum_{(i(v),j(v))\in\Omega(v)}\sum_{v'\neq v}\log P(j(v')|i(v);\Theta) \tag{2.7}$$

En combinant ces different proprieté de de collaboration de données , la méthode développe un embedding unifié qui intègre la diversité intrinsèque à chaque vue et la collaboration entre les vues, tout en prenant en compte les relations de second ordre entre les échantillons.

2.3 Présentation des principaux modèles generant les embbedding multivues

2.3.1 Le modele MvNE (Multi-view Neighbourhood Embedding)

Le modele Multi-view Neighbourhood Embedding (MvNE) représente une approche sophistiquée pour l'intégration unifiée de données multi-vues. Cette méthode se décompose en plusieurs étapes clés.

1. Génération de l'ensemble de données unifié:

   La première étape implique la fusion des différentes vues de l'ensemble de données en une seule représentation unifiée. Chaque vue capte un aspect distinct des données. En cas d'absence de certaines caractéristiques ou échantillons dans des vues, ils seront remplacés par des valeurs nulles. Ainsi, l'ensemble de données unifié englobe l'intégralité des échantillons et caractéristiques issues des diverses vues.

2. Autoencodeur empilé (SAE) pour l'intégration initiale:

   L'autoencodeur empilé (SAE) est un modèle d'apprentissage profond non supervisé. Il sera utilisé sur l'ensemble de données unifié pour generer l'embedding initiale . Le SAE se compose d'un encodeur et d'un décodeur. L'encodeur transforme les données d'entrée en un espace de dimension réduite, formant ainsi l'embedding initial. Et le Decodeur du SAE essayer de reconstitue l'echantillon originale

à partir de sa representation fournit par l'Encodeur . Le modèle SAE sera entraîné en minimisant l'erreur de reconstruction sur la différence entre les données d'entrée x et la sortie reconstruite $\hat{x}$, mesurée par l'erreur quadratique moyenne suivante :

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2 \tag{2.8}$$

3. **Génération de distributions de probabilités unifiées de l'ensemble de données concaténé:**

   Pour chaque échantillon de l'ensemble de données unifié, des probabilités symétriques p_{ij} sont calculées afin d'estimer la probabilité de sélectionner un point voisin. La formule est la suivante :

$$p_{ij} = \frac{\prod_v p_{v_{ij}}}{\prod_v p_{v_{ij}} + \prod_v \sum_{k\neq j} p_{v_{ik}}} \tag{2.9}$$

   - p_{ij} : Probabilité combinée que l'échantillon i choisisse l'échantillon j comme voisin, en tenant compte de toutes les vues.
   - $\prod_v p_{v_{ij}}$: Produit des probabilités $p_{v_{ij}}$ sur toutes les vues v. Chaque $p_{v_{ij}}$ indique la probabilité que, dans la vue v, l'échantillon i choisisse j comme voisin.
   - $\prod_v \sum_{k\neq j} p_{v_{ik}}$: Produit des sommes des probabilités que l'échantillon i choisisse un autre échantillon k (différent de j) comme voisin, calculé pour chaque vue v.

   Les probabilités $p_{v_{ij}}$, basées sur une distribution gaussienne, sont préalablement calculées séparément pour chaque vue. La probabilité est donnée par

$$p_{ij}^{v} = \frac{\exp(-d_{ij}^{v})^2}{\sum_k \exp(-d_{ik}^{v})^2} \tag{2.10}$$

   où d_{ij}^{v} représente la dissimilarité entre les échantillons i et j dans la vue v.

   Cette approche est particulièrement pertinente dans les scénarios nécessitant l'intégration de données provenant de sources diverses pour obtenir une vue complète et unifiée, comme dans l'analyse de données multi-omiques ou la fusion de données issues de capteurs multiples.

4. **Génération de distributions de probabilités dans l'espace intégré:**

   Dans cette étape, nous calculons la probabilité symétrique q_{ij} pour chaque échantillon dans l'espace latent de l'Autoencodeur empilé (SAE). Cette probabilité représente la chance que le point i sélectionne le point j comme voisin. Elle est basée sur une distribution de Student t. L'équation correspondante est

$$q_{ij} = \frac{1 + \|y_i - y_j\|^{2^{-1}}}{\sum_{l\neq k}(1 + \|y_l - y_k\|^2)^{-1}} \tag{2.11}$$

   , où y_i et y_j désignent les représentations des points i et j dans l'espace d'embedding. La norme $\|y_i - y_j\|$ mesure la distance euclidienne entre ces deux points dans cet espace.

Cette formulation permet de capturer les relations de proximité entre les échantillons dans un espace de dimension réduite, facilitant ainsi la compréhension des structures intrinsèques des données multi-vues.

5. **Optimisation de l'intégration unifiée** L'objectif final est de trouver un embedding dans un espace à faible dimension qui reflète au mieux la distribution de probabilité unifiée. Cette optimisation est réalisée en minimisant la divergence de Kullback-Leibler (KL) entre la distribution de probabilité unifiée et celle de l'espace intégré. La formule de divergence KL est exprimée par:

$$C = KL(P||Q) = \sum_{i,j} p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{2.12}$$

La descente de gradient est utilisée pour ajuster itérativement la position des échantillons dans l'espace intégré.

2.3.2 Le modele MANE (Multi-View Collaborative Network Embedding)

Le modèle MANE, développé pour les réseaux multi-vues, part du principe que ces réseaux sont formés de graphes non orientés. Ce modèle s'appuie sur la théorie des modèles de collaboration de données, en examinant de près les interactions et les relations entre les différentes vues. Il utilise la fonction exponentielle pour calculer les probabilités de prédire avec précision la représentation d'un nœud contextuel en fonction des nœuds central. les etapes de la contruction de l'embbeding unifié des noeud est la suivantes :

1. **Construction de l'Ensemble de Paires de Nœuds**

   Cette phase implique la création d'un ensemble de paires de nœuds, divisées en trois catégories :

   (a) Paires de Nœuds Intra-Vue

   Pour générer ce type de paires de nœuds, le modèle définit des marches aléatoires au sein de chaque vue afin de générer des séquences de nœuds. Ces séquences ont pour but de révéler la structure topologique de chaque vue. Ensuite, ces séquences sont tronquées en paires de nœuds

   (b) Paires de Nœuds Inter-Vues Intra-Nœud

   Ces paires sont composées d'instances d'un même nœud dans différentes vues.

   (c) Paires de Nœuds Inter-Vues et Inter-Nœuds

   Incluant des paires d'un nœud dans une vue et de différents nœuds dans une autre, ces associations aident à déchiffrer la collaboration de second ordre.

2. **Definir la dimension de la reperesentation vectorielle des noeuds dans l'espace unifie**

   L'approche MANE consiste à représenter un nœud dans chaque vue, puis à concaténer la représentation de ce nœud dans chaque vue pour former sa représentation finale. L'objectif est de capturer la

diversité propre à chaque vue en traitant les opérations sur les paires de nœuds intra-vue de manière distincte. Pour s'assurer que chaque vue contribue de manière égale à la représentation globale du réseau, le cadre conceptuel de MANE divise l'espace d'embedding entre les différentes vues de manière équitable. Elle définit la fonction de représentation de l'embedding dans une vue de la manière suivante :

$$\begin{cases} f^v : U \to \mathbb{R}^{\lfloor D/|V| \rfloor} \\ \quad i \mapsto f_{i(v)} \end{cases} \tag{2.13}$$

- Ici, U est l'ensemble des nœuds.
- $\mathbb{R}^{\lfloor D/|V| \rfloor}$ est l'ensemble de sortie, représentant l'espace vectoriel dans lequel les embeddings sont placés.
- D est la dimension totale de l'espace d'embedding,
- $|V|$ est le nombre de vues dans le réseau.
- $f_{i(v)}$ represente le vecteur dense du noeud i pour la vue v

3. **Calcul de la fonction de Perte pour l'entrainement du modele**

   (a) **definition de la fontion de perte :**

   pour prendre en compte la diversité intra-vue ainsi que les interactions inter-vues,le framework MANE utilise une combinaison linéaire de fonctions de pertes consue sur les trois typees d'emsembles de noeuds (les trois fonctions de pertes definir dans la theories des modeles de collaboration de données) . elle est de finide la maniere suivante

$$Loss = L_{\text{Div}} + \alpha \cdot L_{C1} + \beta \cdot L_{C2} \tag{2.14}$$

   où :

   - L_{Div} : Représente la perte liée à la diversité intra-vue. Elle vise à capturer la diversité et les caractéristiques uniques de chaque vue individuelle dans le réseau multi-vues.
   - L_{C1} : Correspond à la perte de collaboration de premier ordre. Cette composante de la perte aligne les représentations spécifiques d'un même nœud à travers différentes vues, assurant la cohérence et la similitude des représentations d'un nœud d'une vue à l'autre vue .
   - L_{C2} : Représente la perte de collaboration de second ordre. Elle se concentre sur les relations entre les nœuds à travers différentes vues, exploitant les associations entre les nœuds d'une vue pour renforcer la collaboration entre les différentes vues.
   - α et β : Sont des hyperparamètres qui déterminent l'importance relative des pertes de collaboration de premier et de second ordre par rapport à la perte de diversité. Ces hyperparamètres sont ajustés pour équilibrer le modèle en fonction des particularités du réseau et des objectifs de l'analyse.

(b) **definition de la fonction de probabilité**

La probabilité $P(j(v)|i(v);\Theta)$ est mise en pratique par le modèle MANE via une fonction softmax à travers l'implemetation de l'equation :

$$P(j(v)|i(v);\Theta) = \log\left(\frac{\exp(f_{i(v)} \cdot f_{j(v)})}{\sum_{k\in U}\exp(f_{i(v)} \cdot f_{k(v)})}\right) \tag{2.15}$$

Où :

- $f_{i(v)}$ est la representation vectoreille dense du noeud i dans la vue v
- $f_{i(v')}$ est la representation vectoreille dense du noeud i dans la vue v'
- La fonction exponentielle $\exp(f_{i(v)} \cdot f_{j(v')})$ transforme le score de similarité en probabilité.
- $P(j(v)|i(v);\Theta)$ est la probabilité de prédire correctement la représentation du nœud j dans une vue v', en se basant sur sa connexion avec le nœud i dans la vue v, sous les paramètres du modèle Θ.

Ainsi , nous aurons expresiion des des differeents

$$L_{Div}(\Theta) = -\sum_{v\in V}\sum_{(i(v),j(v))\in\Omega(v)}\log\left(\frac{\exp(f_{i(v)} \cdot f_{j(v)})}{\sum_{k\in U}\exp(f_{i(v)} \cdot f_{k(v)})}\right) \tag{2.16}$$

$$L_{C1}(\Theta) = -\sum_{v\in V}\sum_{(i(v),\cdot)\in\Omega(v)}\sum_{v'\neq v}\log\left(\frac{\exp(f_{i(v)} \cdot f_{i(v')})}{\sum_{k\in U}\exp(f_{k(v)} \cdot f_{i(v')})}\right) \tag{2.17}$$

$$L_{C2}(\Theta) = -\sum_{v\in V}\sum_{(i(v),j(v))\in\Omega(v)}\sum_{v'\neq v}\log\left(\frac{\exp(f_{i(v)} \cdot f_{j(v')})}{\sum_{k\in U}\exp(f_{i(v)} \cdot f_{k(v')})}\right) \tag{2.18}$$

Où :

- V représente l'ensemble des vues dans le réseau multi-vues.
- $\Omega(v)$ désigne les paires de nœuds connectés dans la vue v.
- $i(v)$ et $j(v)$ sont des nœuds connectés dans la vue v, et $j(v')$ est la représentation du nœud j dans une autre vue v'.
- U est l'ensemble total des nœuds dans le réseau.

3

Chapter 3:Méthodologie

Notre recherche, axée sur l'efficacité de la détection de modules actifs dans des réseaux biologiques, adopte une approche innovante basée sur des données multivues. Afin de rester en continuité avec les recherches précédentes, nous envisageons d'intégrer cette approche multivue au sein du framework AMINE, reconnu pour ses performances supérieures dans l'analyse de données univues avec des graphes pondérés au niveau des nœuds. Pour cela, nous établirons un cadre conceptuel qui intègre la construction de différentes vues du graphe et d'autres règles sur le processus d'intégration d'embeddings multivues dans AMINE.

3.1 Cadre conceptuel

Pour atteindre nos objectifs, nous avons défini un cadre conceptuel articulé autour de plusieurs processus clés. Initialement, nous construisons deux vues distinctes à partir d'un graphe de données pondéré, adapté au framework AMINE. Cette approche vise à assurer une continuité avec les recherches précédentes utilisant AMINE, qui a démontré d'excellents résultats dans la détection de modules actifs. La première vue sera influencée par la topologie structurelle du graphe, tandis que la seconde se concentrera sur les poids des nœuds, reflétant les p-values des protéines dans le graphe.

L'étape suivante consiste à unifier ces deux vues dans un espace vectoriel commun. Cette fusion est cruciale pour créer une représentation complète et intégrée des données. Après avoir établi cet espace vectoriel unifié, nous appliquerons l'algorithme glouton d'AMINE sur l'embedding résultant. Cependant, avant d'appliquer cet algorithme, il est essentiel de définir des métriques de similarité adaptées à ce nouvel embedding. Ces métriques seront déterminantes pour évaluer la pertinence et l'efficacité de notre approche multivue dans la détection de modules actifs.

Notre démarche implique également une analyse approfondie des caractéristiques intrinsèques des données biologiques (au niveau de la validation des données) , en tenant compte de la variabilité et de la complexité des interactions génétiques. En intégrant ces aspects, nous visons à améliorer la précision de la détection des modules actifs.

Nous illustrons notre cadre conceptuel avec la Figure 3.1.

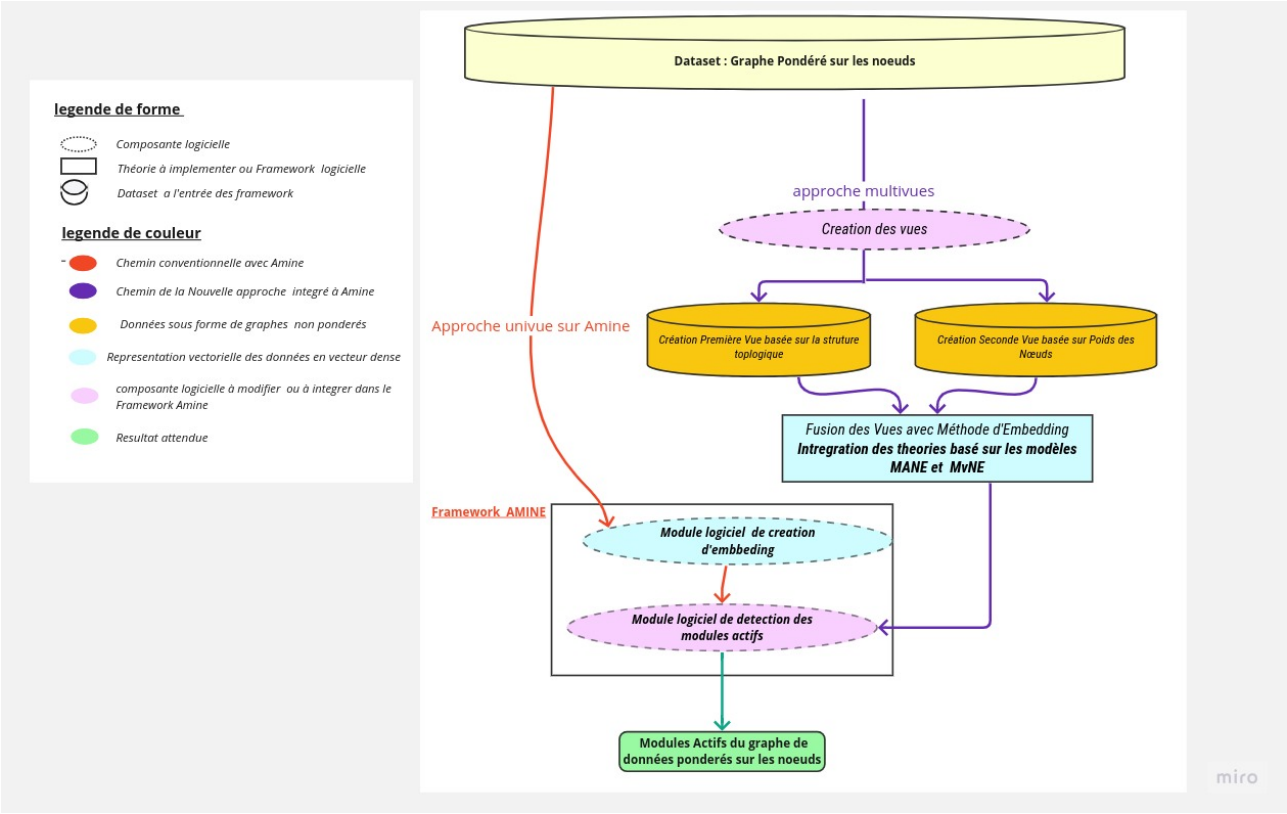

Figure 3.1: Illustration de notre cadre conceptuel via l'approche multivues

3.2 Conception du modèle

3.2.1 Modèles de Construction de la Première Vue

La construction des vues est un élément central de notre recherche. La première vue, en particulier, joue un rôle crucial dans l'analyse des données. Cette vue est conçue comme une réplique fidèle du graphe d'origine, mais sans pondération au niveau des nœuds. En adoptant cette méthode, nous préservons intégralement la structure topologique du graphe, ce qui nous permet de capturer et d'analyser les relations et les connexions intrinsèques entre les différents nœuds. Cette préservation de la structure topologique est essentielle pour deux raisons. Premièrement, elle permet une interprétation plus directe et intuitive des relations entre les nœuds, car chaque lien ou connexion dans le graphe reflète une interaction ou une association réelle, non influencée par des poids. Deuxièmement, en conservant la structure originale, elle nous permet de fidéliser les propriétés de connectivité des protéines pour une activité biologique.

3.2.2 Modèles de Construction de la Deuxième Vue

Pour la deuxième vue, nous avons élaboré quatre modèles de construction distincts, chacun proposant deux sous-variantes basées sur des règles de filtrage spécifiques. Les nœuds qui ne respectent pas ces règles seront traités soit comme des singletons (première variante) soit retirés complètement (seconde variante)

1. **Modèle Construction 1** : Filtrage des composantes connexes du graphe où les relations sont établies

uniquement avec les nœuds ayant une p-value inférieure ou égale à 0.05.

- Sous-Variante 1 :Les autres nœuds non connectés sont considérés comme des singletons.
- Sous-Variante 2 : Ces nœuds sont retirés.

2. **Modèle Construction 2** :Filtrage de tous les nœuds avec une p-value inférieure à 0.05 pour créer une composante connexe complète avec cet ensemble de noeuds de p_value inferieur ou egale à 0.05 .
   - Sous-Variante 1 : Les nœuds avec une p-value supérieure à 0.05 deviennent des singletons.
   - Sous-Variante 2 : Ces nœuds sont retirés.
3. **Modèle Construction 3** : La seconde vue a la même structure topologique que la première, mais avec des arêtes supplémentaires entre les nœuds de p-value inférieure à 0.05 pour former un sous-graphe complet.
   - Sous-Variante 1 : Traitement des nœuds singletons existants.
   - Sous-Variante 2 : dans cette construction la sous variante n'exite pas car dans le graphe d'origine nous n'avons pas de noeud singletions, lors de la construction du graphe nous veillons a ce que toutes les composance connexe soient liés par au moins une arrete .
4. **Modèle Construction 4** : Construction d'un graphe en ajoutant des arêtes entre les nœuds dont la différence de "ZScores" est inférieure ou égale à 0.4.
   - Sous-Variante 1 : Les nœuds non connectés sont considérés comme des singletons.
   - Sous-Variante 2 : Ces nœuds sont retirés

3.2.3 choix du Modèles de theorie d'unifictaion d'embbedding

Le choix entre ces variantes influencera la méthode d'intégration unifiée, en tenant compte des capacités des frameworks comme MANE (Multi-View Collaborative Network Embedding), adapté aux vues complètes, et MvMe (Multi-view Neighbourhood Embedding), plus efficace pour les vues incomplètes. Des tests préliminaires seront réalisés pour évaluer l'efficacité de ces différentes approches afin d'adopter un modèle final.

La sélection du modèle optimal pour la deuxième vue est essentielle, car elle influence directement la qualité des informations utilisées dans notre analyse finale. Chaque modèle d'intergartion unifie sera seront évalués sous les deux variantes de la secondes vues . L'objectif est de déterminer la combinaison la plus efficace pour la détection de modules actifs, en se basant sur des critères tels que la précision, la robustesse et la pertinence biologique, ainsi que la métrique de similarité de validation sur l'embedding unifié. En appliquant ces critères d'évaluation rigoureux, nous nous assurons que le modèle choisi offre non seulement une représentation précise des données biologiques, mais est également robuste et adapté à divers contextes d'analyse.

3.2.4 Critères d'Évaluation

Pour évaluer et valider nos modèles dans le but d'assurer leur efficacié et leur fiabilité , nous avons élaboré trois critères principaux, chacun ciblant un aspect différent de la performance du modèle :

1. **la Métrique Précision :**

   La capacité du modèle à identifier correctement les modules actifs est mesurée par le score F1. Ce critère évalue l'équilibre entre la précision (proportion de vrais positifs parmi les identifications) et le rappel (proportion de vrais positifs parmi les cas réels). La formule du score F1 est la suivante :

   $$\text{F1 Score} = 2 \times \frac{\text{Précision} \times \text{Rappel}}{\text{Précision} + \text{Rappel}} \tag{3.1}$$

   Où :

   - La précision est calculée comme $\frac{TP}{TP+FP}$
   - Le rappel est calculé comme $\frac{TP}{TP+FN}$
   - TP représente le nombre de vrais positifs (Vrai positives).
   - FP représente le nombre de faux positifs (Faux positives).
   - FN représente le nombre de faux négatifs (Faux negatives).

2. **Métrique de similarité:** Nous avons défini trois métriques dans l'espace unifié pour évaluer la similarité des représentations vectorielles des nœuds. Ces métriques comprennent la similarité cosinus, la distance euclidienne, et la similarité de Pearson.

   - ***La Similarité de Pearson :*** La similarité de Pearson (ou corrélation de Pearson) mesure la corrélation linéaire entre deux variables aléatoires. La formule pour calculer la corrélation de Pearson entre deux vecteurs X et Y est :

     $$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \tag{3.2}$$

     où $\bar{X}$ et $\bar{Y}$ sont les moyennes des vecteurs X et Y respectivement. La valeur de r se situe entre -1 et 1.

   - ***La Similarité Cosinus :***

     La similarité cosinus mesure l'angle entre deux vecteurs dans un espace vectoriel. Elle est calculée comme le produit scalaire des vecteurs normalisé par le produit de leurs normes. La formule est :

     $$\text{similarity\_cosine} = \frac{X \cdot Y}{\|X\|\|Y\|} \tag{3.3}$$

     où $X \cdot Y$ est le produit scalaire des vecteurs X et Y, et $\|X\|$ et $\|Y\|$ sont les normes de X et Y.

- ***La Similarité Euclidienne :***

  La similarité euclidienne (ou distance euclidienne normalisée) est basée sur la distance euclidienne entre deux points. Si D est la distance euclidienne entre deux vecteurs X et Y, la similarité S est calculée comme :

$$\text{Similarity\_euclidienne} = \frac{1}{1+D} \tag{3.4}$$

  où D est la distance euclidienne, calculée comme :

$$D = \sqrt{\sum_{i=1}^{n} (X_i - Y_i)^2}$$

  Chacune de ces métriques de similarité apporte une perspective différente sur la manière dont les nœuds ou les modules sont liés ou distants les uns des autres dans l'espace vectoriel.

3. **Robustesse** (Test sur Ensemble de Données Artificiel):

   La stabilité du modèle est testée sur un ensemble de données artificiel comprenant 1000 graphes. Ce critère évalue la capacité du modèle à maintenir une performance constante malgré les variations dans les données d'entrée, assurant ainsi sa robustesse et sa fiabilité.

Nous pouvons illustrons notre modele conceptuel avec la Figure 3.2.

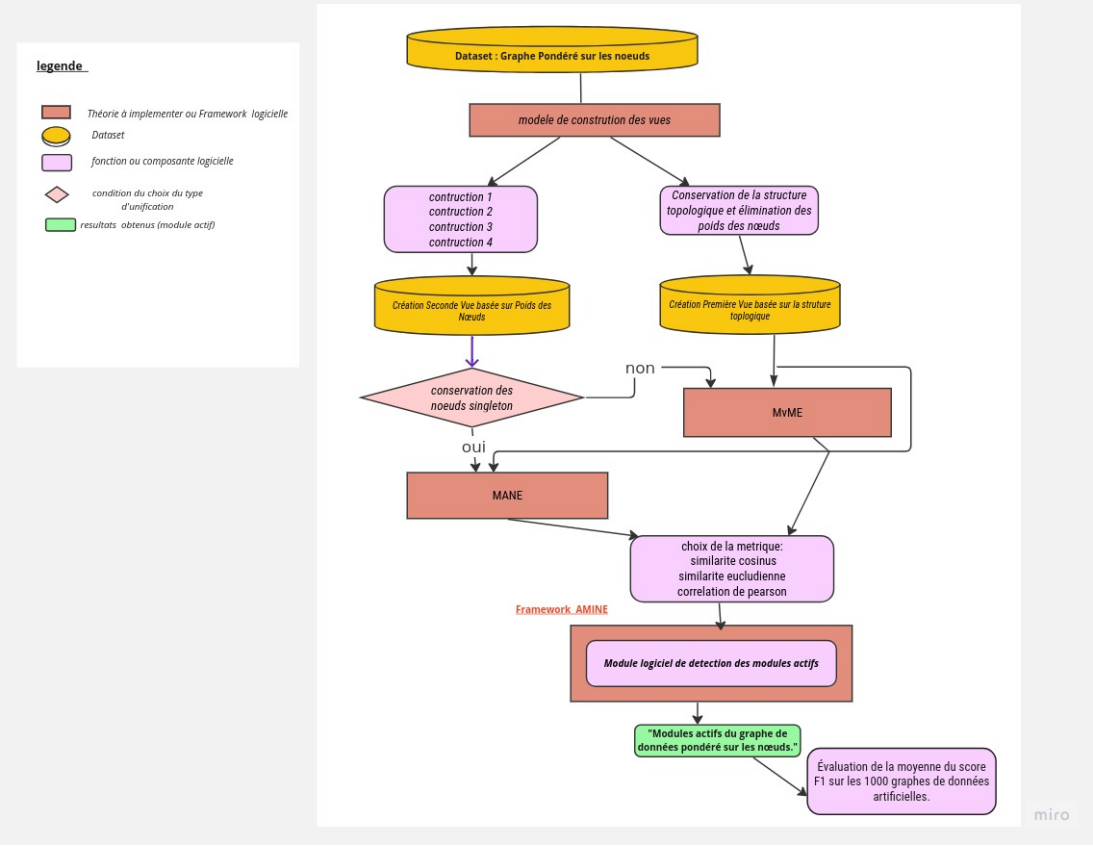

Figure 3.2: Illustration de notre modele conceptuel

3.3 Collecte de données:

Dans le cadre de notre étude, nous avons opté pour la création de données artificielles afin de simuler des réseaux biologiques complexes. Cette méthode nous permet de contrôler précisément les paramètres du réseau, ce qui est crucial pour tester l'efficacité de notre modèle de détection de modules actifs. Après avoir validé le modèle sur ces données artificielles, nous l'appliquerons sur des données réelles, celles générées dans l'article de Chiou et al. Le modèle AMINE a également été évalué sur ces données.

3.3.1 Description de l'Algorithme de Génération de Données

Cet algorithme génère des données en simulant des réseaux biologiques complexes à l'aide de graphes sans échelle. Il s'appuie sur une version améliorée de la méthode de Barabási-Albert, idéale pour modéliser des réseaux où la distribution des degrés des nœuds suit une loi de puissance. En plus d'intégrer la méthode de Barabási-Albert, l'algorithme comprend trois fonctions principales qui sont utiles dans la construction des clusters :

1. **neighbors** : son Rôle est de détermine le nombre de voisins d'un nœud spécifique à une distance d'orde k. Partant d'un nœud initial ($start$), elle explore et compte les voisins jusqu'à atteindre le niveau k, offrant ainsi une vue sur la connectivité locale du nœud dans le graphe.

2. **knbrs** : Cette fonction identifie tous les voisins d'un nœud à un niveau k. Similaire à *neighbors_order*, elle retourne un ensemble des voisins jusqu'au niveau k, permettant de comprendre les interactions potentielles d'un nœud donné.

3. **get_seeds** : Cette fonction sélectionne des nœuds initiaux pour la création de modules dans le graphe. et veuille a ce que les nœuds soient suffisamment éloignés les uns des autres ($min\_distance$), assurant ainsi une distribution équilibrée et non chevauchante des modules dans le graphe.

Chaque sous-fonction joue un rôle clé dans l'élaboration d'un graphe complexe et structuré, reflétant les propriétés des réseaux biologiques. *neighbors_order* et *knbrs* sont cruciales pour analyser la structure locale des nœuds, tandis que *get_seeds* est essentielle pour initier la formation de modules distincts au sein du graphe.

3.3.2 Génération de la structure topologique du graphe

la structure topologique du reseaux de données est basé sur le modèle de Barabási-Albert étendu, permettant la création de graphes libres d'échelle avec la propriete fondamentale de l'attachement préférentiel (Les nouveaus noeuds ont tendance à se connecter à des nœuds déjà bien connecté). Les paramètres clés de ce modele sont :

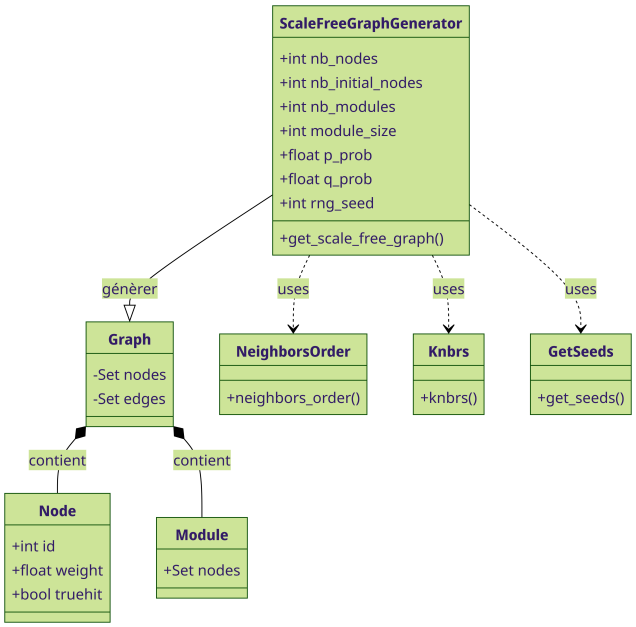

Figure 3.3: Illustration en digaramme classe de la desciption de l"agorithme qui genere les données artificielles

1. Nœuds Initiaux (*nb_initial_nodes* =*3*) : Les nœuds initiaux forment le noyau de départ du graphe. Ils sont essentiels pour commencer le processus de croissance du réseau selon la méthode de Barabási-Albert. Le nombre de nœuds initiaux influence la structure initiale du réseau. Un petit nombre de nœuds initiaux peut conduire à un réseau plus centralisé autour de ces nœuds, tandis qu'un plus grand nombre peut favoriser une structure plus distribuée. Le choix du nombre de nœuds initiaux doit refléter l'objectif de la simulation. Pour un réseau biologique, il est souvent souhaitable de commencer avec un petit nombre de nœuds initiaux pour simuler le développement naturel d'un réseau biologique à partir de quelques éléments clés.

2. Probabilités (p_prob) et (q_prob) : la probabilité (p_prob=0.09): Contrôle l'ajout de nouvelles arêtes entre les nœuds existants. Un (p_prob) élevé favorise la création de nouvelles connexions. et la probabilité (q_prob=0.7): Gère la réorganisation des arêtes existantes. Un (q_prob) élevé permet une plus grande dynamique dans la structure du réseau. pour assure un équilibre entre la croissance et la reorganistaion du réseau la somme de probabilité doit soumise a la contrainte suivante (p_prob) + (q_prob) < 1,

Après avoir construit la structure topologique du graphe basée sur le modèle de Barabási-Albert, nous veillerons à ne pas laisser de composantes connexes disjointes. Le principe est simple,il s'agira de créer un lien aléatoire entre les composantes connexes.

3.3.3 Génération des Modules dans le Graphe

La formation des modules dans le réseau est une étape cruciale, simulant la création de groupes de gènes ou de protéines fonctionnellement liés. Des nœuds "graines" sont sélectionnés en fonction de leur degré

de connectivité et de leur distance relative, assurant une répartition équilibrée des modules dans le graphe. Autour de chaque graine, un module est formé en intégrant des nœuds voisins, choisis selon un processus aléatoire pondéré par la distance dans le graphe. La taille de chaque module est contrôlée par le paramètre 'module_size'. Par exemple, si une graine est sélectionnée, les nœuds à une distance de 1 ou 2 pas sont progressivement inclus dans le module, en fonction de leur probabilité de connexion.

3.3.4 Attribution des Poids (p_value) aux Nœuds

Enfin, des poids sont attribués à chaque nœud du graphe pour simuler des caractéristiques biologiques spécifiques. Les nœuds hors modules reçoivent des poids selon une distribution uniforme [0,1], tandis que ceux au sein des modules suivent une distribution normale tronquée. Cette distribution est choisie pour refléter une concentration élevée de caractéristiques biologiquement significatives dans les modules, comme on pourrait s'y attendre dans des groupes de gènes ou de protéines actifs. Les poids des nœuds dans les modules sont donc générés selon la formule :

$$P(\text{p\_value}) = \text{TruncNorm}(\mu, \sigma, a, b) \tag{3.5}$$

où $\mu = 0$, $\sigma = 0.05$, et les bornes a et b sont ajustées pour maintenir les poids entre 0 et 1.

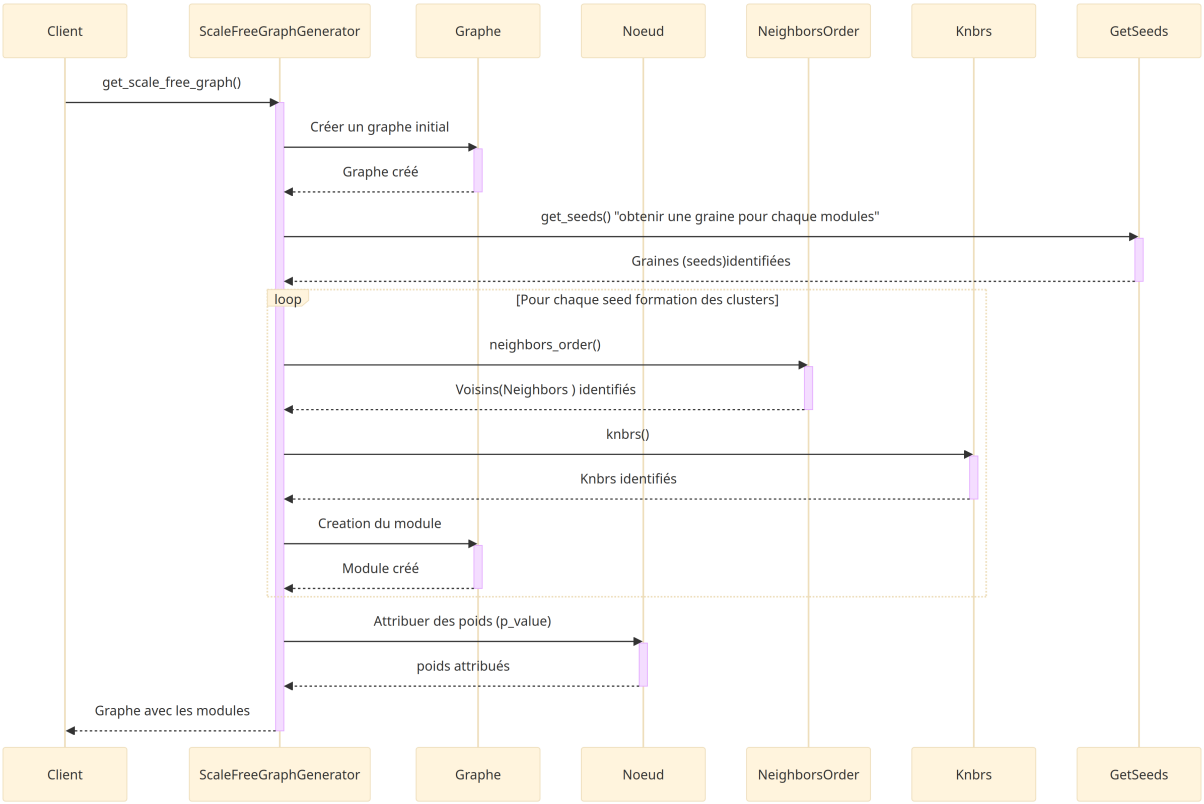

Figure 3.4: Illustration la gereration des données artificielles

La méthodologie proposée offre une nouvelle perspective pour l'analyse des interactions complexes au sein des réseaux biologiques en intégrant le concept de données multivues et en appliquant des critères

d’évaluation rigoureux, nous visons à améliorer la précision et la robustesse de la détection des modules actifs, un aspect crucial pour la compréhension des mécanismes biologiques sous-jacents

4

Chapter 4: Validation du Modèle

4.1 Approche de simulation

4.2 Justification des méthodes choisies

4.3 Statistiques descriptives

Modèle	Mean	Variance	Q25	Q50 Médiane	Q75	Moy. sur 81 graphes	Moy. sur 107 gra
Amine	0.54214	0.27516	0.40000	0.63397	0.73684	0.59725	0.59038
Mane Pearson	0.34995	0.12707	0.27273	0.35294	0.43243	0.33093	0.34313
Mane Euclidien	0.30884	0.13152	0.21525	0.30303	0.41667	0.30553	
Mane Cosinus	0.33250	0.13056	0.25532	0.33333	0.42105	0.33250	

Table 1: Analyse comparative des modèles

Modèle	Mean	Variance	Q25	Q50 Médian	Q75	Moy. sur 42 graphes	Moy. sur 86 graphes
Amine	0.5486	0.1861	0.4666	0.5806	0.6666	0.49442	0.52324
Mane Pearson	0.4824	0.1083	0.4084	0.4846	0.5614	0.44970	0.48010
Mane Euclidien	0.4546	0.0940	0.4106	0.4404	0.4908	0.45463	
Mane Cosinus	0.4856	0.1080	0.4154	0.4807	0.5696	0.45329	0.48562

Table 2: Analyse comparative des modèles pour 20_G1

Modèle	Mean	Variance	Q25	Q50 Médiane	Q75	Moy. sur 112 graphes	Moy. sur 332 gra
Amine	0.5486	0.1861	0.4666	0.5806	0.6666	0.52640	0.53010
Mane Pearson	0.3405	0.0531	0.30380	0.33766	0.37209	0.34055	
Mane Euclidien	0.3380	0.0559	0.30000	0.33735	0.37333	0.35133	0.34423
Mane Cosinus	0.3372	0.0530	0.29971	0.33333	0.37397	0.34081	

Table 3: Analyse comparative des modèles hhhhhhpour 20_G2

4.4 Statistiques inférentielles

4.5 Réponses aux questions de recherche

4.6 Visualisation

4.7 Analyse de sensibilité des cluster

4.8 Comparaison avec d'autres modèles clusters

5 Discussion

5.1 Comparaison avec Amine

5.2 Interprétation des résultats

5.3 Implications pratiques

5.4 Limites de l'étude

5.5 Contributions de l'étude

5.6 Suggestions pour des travaux futurs

6 Conclusion

6.1 Résumé des résultats

6.2 Réponses aux questions de recherche

6.3 Contributions de l'étude

6.4 Limites de l'étude

6.5 Suggestions pour des travaux futurs

6.6 Réflexion personnelle

Bibliographie

[Iri17] Irina Gaynanova, Gen Li (2017). “Structural learning and integrative decomposition of multi-view data”. In: *Journal of Statistical Software*. URL: https://arxiv.org/pdf/1707.06573.pdf.

[Loc+13] Lock, Eric F. et al. (2013). “Discovering Structure in High-Dimensional Data Through Methodology and Application of the Joint and Individual Variation Explained (JIVE) Approach”. In: *Journal of Multivariate Analysis*. URL: https://arxiv.org/pdf/1707.06573.pdf.

[OS18] Okuno, Akifumi and Tetsuya Hada Hidetoshi Shimodaira (Feb. 2018). “A probabilistic framework for multi-view feature learning with many-to-many associations via neural networks”. In: DOI: 10.48550/arXiv.1802.04630. URL: https://arxiv.org/abs/1802.04630.

[Pas+22] Pasquier, Claude et al. (June 2022). “A network embedding approach to identify active modules in biological interaction networks”. In: *Life Science Alliance*. DOI: 10.26508/lsa.202201550. URL: https://www.life-science-alliance.org/content/6/9/e202201550.

[Rap+07] Rapaport, P. et al. (2007). “Classification of microarray data using gene networks”. In: *BMC Bioinformatics* 8. DOI: 10.1186/1471-2105-8-35. URL: https://arxiv.org/pdf/1707.06573.pdf.

[XTX13] Xu, Xinxing, Ivor W. Tsang, and Dong Xu (2013). “Soft Margin Multiple Kernel Learning”. In: *Journal of Multivariate Analysis*. DOI: 10.1109/TNNLS.2012.2237183.