République Algérienne Démocratique et Populaire
Ministère de l'enseignement supérieur et de la recherche scientifique
UNIVERSITE FERHAT ABBAS – SETIF-1

Faculté des sciences
Département d'informatique

THESE

Présentée en vue de l'obtention du diplôme de Doctorat en Sciences

Apprentissage et extraction des connaissances dans les bases de données spatiotemporelles

Domaine : Mathématiques et Informatique
Filière : Informatique
Spécialité : Ingénierie des Systèmes Informatique

Par :
MEKROUD Noureddine

Directeur de Thèse :

Pr. MOUSSAOUI Abdelouahab	Professeur	Université Ferhat Abbas, Sétif-1

DEVANT LE JURY

Président :

Pr. BOUAMAMA Salim	Professeur	Université Ferhat Abbas, Sétif-1

Examinateurs :

Pr. CHIKHI Salim	Professeur	Université Constantine-2
Dr. BOUZENADA Mourad	MCA	Université Constantine-2
Dr. MEHENNI Tahar	MCA	Université Msila

Invité :

Dr. Rachid HEDJAM	PHD	Bishop's University, CANADA

Soutenue le : Mercredi 15 avril 2026

Dédicaces

A ceux qui me sont chers…

Au peuple Palestinien

Remerciements

Avant tous, louanges à *Dieu* le tout puissant et miséricordieux, pour m'avoir donné la volonté et la capacité de réaliser ce travail.

A mes chers parents, à ma chère grande et petite famille.

Aussi, je souhaite remercier vivement le professeur MOUSSAOUI Abdelouahab pour ses encouragements et orientations le long de mon parcours de recherche.

Je remercie le professeur Salim BOUAMAMA pour m'avoir fait l'honneur de présider mon jury de thèse. Je remercie également le professeur CHIKHI Salim et le docteur BOUZENADA Mourad et le docteur MEHENNI Tahar pour l'intérêt qu'ils ont bien voulu porter à mon travail en acceptant la lourde tâche d'examinateur. Ravi de voir mon vieux camarade et ami le docteur HEDJAM Rachid comme invité dans cette soutenance. Merci messieurs pour votre présence.

Je réserve les derniers remerciements à tous mes collègues et amis, spécialement au docteur BELBAL Samir, merci pour votre encouragement.

ملخص :
يعد النقص سمة مشتركة في اغلب البيانات الصادرة من الواقع ، على الرغم من أنه عادةً ما تخفي هذه البيانات معرفة ذات أهمية كبيرة. إن استخدام نظريات عدم اليقين لنمذجة مناطق التعبير الجيني (Gen_Exp) في الجنين سيساعد على استخلاص العلاقات الخفية بين الجينات، مع الأخذ في الاعتبار احتمال نقص الدقة في حدود مناطق التعبير الجيني و في تدرج قوة هذه التعبيرات. في هذه الاطروحة، نقترح نمذجة بمنطق الحدود متداخلة و منطق الإمكانية للبيانات الزمانية-المكانية الخاصة بتسلسل صور التعبير الجيني ISH (In Situ Hybridization) التي تمثل مناطق تعبير الجينات عبر مختلف مراحل نمو الجنين للكائن النموذجي" فأرة Edinburgh " .بعد سلسلة من خطوات المعالجة المسبقة لهذه الصور لتحسين استخراج الميزات، نقترح تكييف خوارزمية Apriori مع منطق الحدود متداخلة و منطق الإمكانية لاستخراج عدة انواع من علاقات الارتباط (Association Rules) ، والتي تمثل الارتباطات المكانية بين مناطق التعبير الجيني في الجنين، و في منظور المزدوج تمثل الارتباطات الزمنية وكذلك العلاقات بين الجينات التي تشترك في التعبير في تسلسلات صورISH . أخيرًا، سيتم تقديم تحليل لعلاقات الارتباط المستخرجة بمنطق الحدود متداخلة و منطق الإمكانية، لتحديد منطق النمذجة الأكثر ملاءمة لهذه البيانات. كما أن التفسير البيولوجي للنتائج التي تم الحصول عليها يؤكد ملاءمتها لمبادئ المجال. ستساعد المعرفة المستخرجة علماء الأحياء على فهم التفاعلات بين الجينات بشكل أفضل، لاكتشاف مجموعات الجينات المعبر عنها بشكل مشترك والتي لها نفس الدور الوظيفي، ونمذجة التعبير الجيني الطبيعي لاكتشاف التعبير الجيني غير الطبيعي الذي يمكن أن يسبب أمراضًا وراثية.

الكلمات المفتاحية: علاقات الارتباط ؛ البيانات الزمانية-المكانية. صور ISH ؛ التعبير الجيني. منطق الحدود متداخلة و منطق الإمكانية

Abstract

Imperfection is a common feature in almost all real-world data, although it usually hides crucial knowledge of major interest. Using theories of uncertainty to model gene expression (Gen_Exp) areas in the embryo will help to extract hidden relationships between genes, while taking into account possible imprecision in the boundaries of Gen_Exp zones and the nuanced strength of these expressions. In this paper, we propose a fuzzy and possibilistic modeling of spatiotemporal data from In Situ Hybridization (ISH) sequences of images representing Gen_Exp areas in different embryonic developmental phases of the model species Edinburgh Mouse. Following a series of pre-processing steps on these images to improve feature extraction, we propose an adaptation of the Apriori algorithm to the fuzzy and possibilistic logic for mining two types of Association Rules (AR), which will represent in primal the spatial correlations between Gen_Exp areas in the embryo, and in dual the temporal correlations as well as the relationships between genes that co-express in these ISH image sequences. Finally, an analysis of extracted spatiotemporal fuzzy and possibilistic AR will be provided to decide which modeling is the most suitable. Also, biological interpretation of the obtained results confirms their adequacy with domain principles. The extracted knowledge will help biologists to better understand the interactions between genes, for discovering the co-expressed sets of genes having the same functional role, and modeling normal Gen_Exp for detecting abnormal Gen_Exp that can cause genetic diseases.

Keywords : Association rules; Spatiotemporal data; ISH images; Gene expression; Fuzzy and Possibilistic logic

Résumé

L'imperfection est une caractéristique commune dans presque toutes les données du monde réel, même si elles cachent généralement des connaissances cruciales et d'un intérêt majeur. L'utilisation de théories de l'incertitude pour modéliser les zones d'expression génique (Exp_Gen) dans les embryons permettra d'extraire les relations cachées entre les gènes, tout en tenant compte de l'imprécision possible des limites de ces zones et de la force nuancée de ces expressions. Dans cette thèse, nous proposons une modélisation floue et possibiliste des données spatio-temporelles issues de séquences d'images ISH (In Situ Hybridization) représentant les zones d'Exp_Gen à différentes phases de développement embryonnaire de l'espèce modèle *souris d'Edinburgh*. Après une série d'étapes de prétraitement sur ces images pour améliorer l'extraction des caractéristiques, nous proposons une adaptation de l'algorithme Apriori à la logique floue et possibiliste pour l'extraction de deux types de règles d'association (RA), qui représenteront en Primal les relations spatiales entre les zones d'Exp_Gen dans l'embryon, et en Dual les relations temporelles ainsi que les relations entre les gènes qui co-expriment dans ces séquences d'images ISH. Enfin, une analyse des AR spatiotemporelles floues et possibilistes extraites sera fournie afin de déterminer la modélisation la plus adéquate. De plus, l'interprétation biologique des résultats obtenus confirme leur adéquation aux principes de la biologie génétique. Les connaissances extraites aideront les biologistes à mieux comprendre les interactions entre gènes, à identifier les ensembles de gènes qui co-expriment et ayant le même rôle fonctionnel, et enfin à modéliser l'expression génétique normale pour détecter les expressions génétiques anormales susceptibles de provoquer des maladies génétiques.

Mots-clés : Règles d'association ; Données spatiotemporelles ; Images ISH ; Expression génétique ; Logique floue et possibiliste

TABLE DES MATIERES

LISTE DES FIGURES

LISTE DES TABLEAUX

LISTE DES ABREVIATIONS

RA ..... Règles d'Association
Exp_Gen ..... Expression Génétique
ECD ..... Extraction des Connaissances à partir des Données
DM ..... Data Mining
ISH ..... In Situ Hybridization
ADN ..... Acide Désoxyribo-Nucléique
ARN ..... Acide Ribo-Nucléique
ARNseq ..... ARN Sequencing
EMAGE ..... Edinburgh Mouse Atlas of Gene Expression
RVB ..... Rouge Vert Bleu
PNG ..... Portable Network Graphics
TS ..... Theiler Stages
SEF ..... Sous Ensemble Flou
Fonc_App ..... Fonction d'Appartenance
Deg_App ..... Degré d'Appartenance
Supp ..... Support
Conf ..... Confiance
Min_Sup ..... Support Minimal
Min_Conf ..... Confiance Minimales
Fuzzy_Sup ..... Support Flou
Fuzzy_Conf ..... Confiance Floue
Poss_Sup ..... Support Possibiliste
Poss_Conf ..... Confiance Possibiliste
MAP ..... Maximum A Posteriori
STRING-DB ..... Search Tool for the Retrieval of INteracting Genes/proteins - DataBase

I. INTRODUCTION GENERALE

L'augmentation phénoménale du volume des données spatio-temporelles produites dans différents domaines de recherche a engendré divers problèmes de modélisation et d'interprétation de ces données complexes, qui dissimulent une grande quantité de connaissances cruciales.

La bioinformatique, un domaine de recherche en plein essor, requiert des méthodes avancées d'analyse pour traiter les données biologiques qui sont hétérogènes et vitales, en pleine expansion, mais souvent imprécises et/ou incomplètes, et dont l'interprétation reste un défi. Les données d'expression génétique intègrent des connaissances très riches sur les stades de développement embryonnaire de toute espèce modèle, mais (en grande partie) ces connaissances restent inconnues et donc inexploitées par les biologistes. Parmi les puissantes formes de représentation des expressions génétiques figurent les séquences d'images ISH (*In Situ Hybridization*), qui indiquent in situ (sur les images embryonnaires des organismes vivants) les zones d'expression de chaque gène au cours des phases de croissance de l'embryon de certaines espèces typiques. Etudiées essentiellement pour la détection de gènes qui co-expriment lors des phases de développement des espèces modèles, ces images sont une importante source de données biologiques spatio-temporelles, et leur analyse aide à la pré-détection de maladies d'origine génétique durant les phases préliminaires de développement des espèces vivantes.

Problèmatique

Les règles d'association (RA) sont l'un des algorithmes les plus utilisés en apprentissage automatique. Elles permettent d'extraire des relations cachées entre un ensemble d'attributs initialement représentés sous forme binaire. Mais, lors de l'étude des datasets issus du monde réel, les chercheurs traitent généralement des données ambiguës sur toute situation, soit parce qu'ils doutent de leur véracité (c'est l'*incertitude*), ou parce qu'ils ont des difficultés à les énoncer clairement (c'est l'*imprécision*). Par conséquent, la représentation en format binaire strict (vrai ou faux) des données réelles entraîne généralement une perte importante de précision dans leur quantification et leur sémantique. Ainsi, d'autres versions des RA ont été proposées pour se rapprocher de plus en plus du raisonnement humain (qui est généralement nuancé et loin d'être binaire) et de la réalité des données analysées, cela en adaptant cet algorithme pour être en adéquation avec la nature de ces données incertaines et/ou imprécises, via des modélisations en logique floue, possibiliste, ou en logique évidentielle. Ces adaptations permettent d'extraire des connaissances plus pertinentes et fidèles à la réalité des données étudiées.

L'extraction des RA est aussi une technique efficace pour extraire les relations cachées dans les données d'expression génétique (Exp_Gen), notamment pour représenter formellement les co-expressions génétiques en détectant les RA les plus intéressantes entre gènes, en fonction

de mesures d'intérêts objectifs (intérêts purement computationnels) ou subjectifs (RA biologiquement interprétables).

Dans la littérature existante, la modélisation des données ISH d'expression génétique repose exclusivement sur une représentation binaire. Une telle simplification peut nuire à la qualité des connaissances extraites, car elle suppose que les données sont précises et exemptes d'ambiguïtés. Or, en pratique, ces données peuvent comporter des incertitudes, puisque par exemple l'expression d'un gène dans une zone donnée, à un stade précis et avec un niveau d'intensité donné n'est jamais absolument certaine, ainsi que des imprécisions, notamment dans la délimitation exacte des régions d'expression. Ignorer ces aspects peut conduire à des résultats biologiques moins pertinents et moins fiables.

Ainsi, peu d'efforts de recherche ont été fournis concernant l'extraction de RA floues à partir de données d'Exp_Gen, et aucune modélisation des RA en logique floue n'a été proposée pour l'analyse des données de type ISH, qui sont par nature incertaines et imprécises. De plus, aucun travail scientifique (à notre connaissance) n'a étudié ce type de données d'Exp_Gen en les modélisant via une logique possibiliste, qui est une vision plus large de la logique floue, pour les analyser via les algorithmes d'intelligence artificielle, comme les RA.

Objectifs

Dans cette perspective, notre travail vise à proposer une étude fidèle à la réalité biologique des images ISH d'Exp_Gen, résultant du suivi des stades de développement de l'espèce modèle « Souris d'Edinburgh ». Nous nous intéressons à l'adaptation de l'algorithme Apriori aux théories de l'incertain pour modéliser les données analysées afin d'extraire plusieurs types de RA utiles, suivant deux logiques de modélisation : floue et possibiliste. De plus, le niveau nuancé d'Exp_Gen (fort, modéré et non détecté) sera pris en compte pour extraire les caractéristiques les plus intéressantes représentant la réalité de ces Exp_Gen spatio-temporelles.

Après une série d'opérations de prétraitement et d'extraction de caractéristiques visant à réduire la dimensionnalité des images étudiées (et donc la complexité algorithmique) tout en préservant leur variabilité, nous cherchons à extraire : les relations entre les Exp_Gen en détectant les ensembles de gènes qui co-expriment, les régions dont l'expression est synchronisée et les relations temporelles entre les gènes qui co-expriment à différents stades du développement embryonnaire. Les RA extraites permettent de découvrir une variété de connaissances biologiques cruciales, cachées dans ces images ISH.

L'adaptation proposée de l'algorithme Apriori (initialement conçu pour les données binaires) repose sur une redéfinition des notions de *cardinalité* et d'*union* entre les ensembles d'attributs (itemsets). De plus, nous utilisons une méthode basée sur les *normes triangulaires* pour calculer le support des itemsets et la confiance des règles extraites, afin de les rendre compatibles avec les logiques floues et possibilistes.

Enfin, une analyse comparative de la qualité des règles extraites suivant les deux modélisations (floue et possibiliste) sera proposée, afin de déterminer la modélisation la plus

adéquate aux données biologiques étudiées. De plus, une analyse fonctionnelle des gènes formant les itemsets des RA extraites montre qu'une grande partie des co-expressions géniques détectées est conforme aux principes biologiques, ce qui confirme que ces gènes interagissent effectivement lors de la formation de différents systèmes et organes biologiques.

Contenu du document

Ce document est organisé en quatre chapitres, chacun réparti en deux parties comme suit :

- **Le chapitre 1 :** la première partie présente une introduction générale à la bio-informatique et à ses principales banques de données ; la seconde partie est consacrée aux données d'expression génétique.
- **Le chapitre 2** : la première partie décrit les principes de l'Extraction de Connaissances à partir des Données (ECD) et les principaux algorithmes de *Machine Learning* ; la seconde partie détaille l'algorithme de règles d'association utilisé dans notre approche.
- **Le chapitre 3** : sa première partie décrit les bases théoriques des différentes logiques de modélisation des données incertaines ; en deuxième partie on présente les adaptations nécessaires de l'algorithme Apriori aux logiques de l'incertain.
- **Le chapitre 4** : la première partie explique l'approche proposée via une présentation détaillée des deux modélisations proposées (basées sur la logique floue et possibiliste). La seconde partie présente les aspects techniques de l'implémentation, suivis par la représentation et l'interprétation biologique des résultats obtenus, puis une analyse et discussion sur la pertinence des connaissances biologiques extraites et leur utilité dans la compréhension des interactions entre gènes.

Le document finira par une conclusion générale et des perspectives de recherche futures.

Chapitre 1

Chapitre I : La Bio-Informatique ET Les images d'expression génétique

I.1 Introduction

La biologie vise l'étude des êtres vivants, en se focalisant sur des espèces modèles. En raison de la grande quantité de données biologiques générées par le séquençage des génomes d'organismes et l'importance des connaissances qu'elles cachent, la bio-informatique, comme discipline consistant à concevoir des outils informatiques pour analyser des données biologiques, connaît d'énormes progrès dans la puissance de traitement et la pertinence des connaissances découvertes.

Dans ce chapitre on va présenter les notions de base de la biologie et l'historique de la bio-informatique et ses principaux domaines d'application. On présentera aussi les divers types de banques de données biologiques, en se focalisant dans la deuxième partie de ce chapitre sur les données d'expression génétique, avec un intérêt particulier porté aux images d'hybridation *in situ* (ISH) de l'espèce modèle *Edinburgh Mouse*.

Partie 1 : La Bio-Informatique

I.2 Notions biologiques de base

I.2.1 Les diverses définitions de la Bio-informatique

La Bio-informatique est un champ de recherche multidisciplinaire récent qui implique la biologie, l'informatique et les mathématiques ; dont l'objectif est d'analyser automatiquement et de produire des connaissances nouvelles à partir des données biologiques qui sont généralement volumineuses et de différents types : séquences d'ADN, structures protéiques, images d'expressions génétiques ...etc [1].

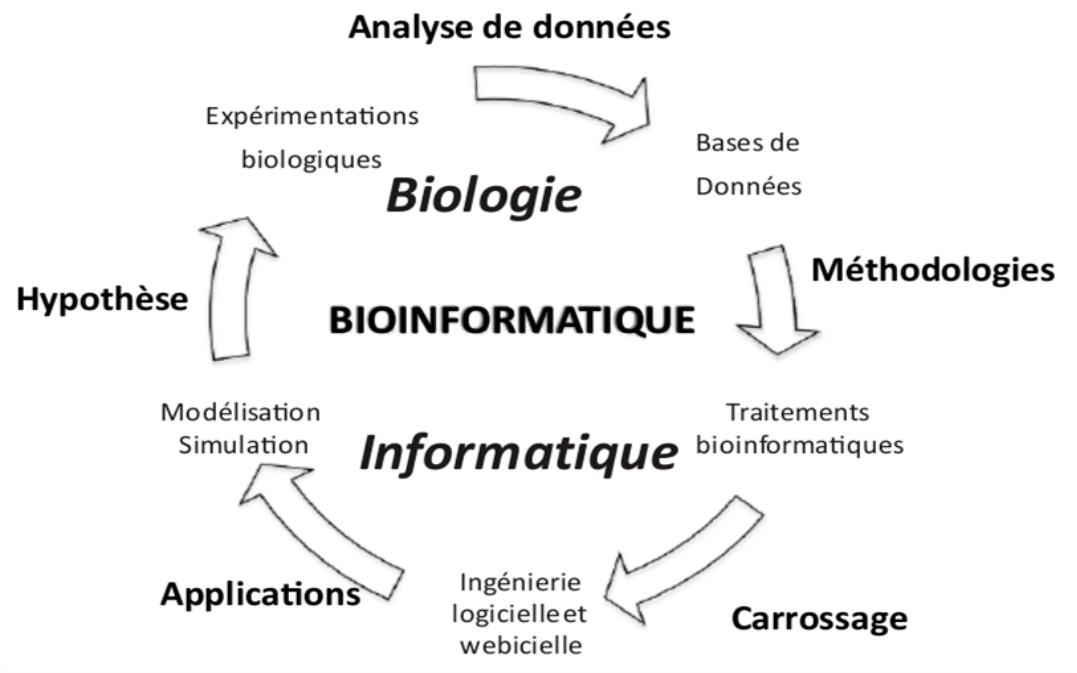


Figure I.1 : L'interaction entre la biologie et l'informatique [2]

Le terme bio-informatique a été inventé par Paulien Hogeweg et Ben Hesper de l'université d'Utrecht, Pays-Bas dans les années 70 pour décrire "l'étude des processus informatiques dans les systèmes biotiques". Ce terme a été utilisé tôt lorsque les premières données de séquences biologiques ont commencé à être partagées. Alors que les premières méthodes d'analyse n'étaient pas assez sophistiquées, la bio-informatique est aujourd'hui considérée comme une discipline beaucoup plus large, englobant la modélisation et l'analyse d'images, en plus des méthodes classiques utilisées pour la comparaison de séquences d'ADN linéaires et des structures protéiques tridimensionnelles [1].

La révolution biologique a été initiée par quatre événements majeurs : la découverte de l'ADN en tant que support de l'information génétique (1944) et de sa structure en double hélice (1953), ainsi que la mise en place du dogme central de la biologie moléculaire (1958) et le déchiffrage du code génétique (1962). Ces événements constituent les fondements de la génomique. Ensuite, à partir de la fin des années 1980, le terme « bio-informatique » a principalement été utilisé pour désigner des méthodes computationnelles permettant une analyse automatisée des données du génome des êtres vivants [3].

En littérale, on peut trouver plusieurs définitions de la bio-informatique :

Définition 1 : D'après *the National Human Genome Research Institute* (NHGRI), « la bio-informatique est la branche de la biologie qui s'occupe de l'acquisition, du stockage, de l'affichage et de l'analyse des informations trouvées dans les données de séquences d'acides nucléiques et de protéines », l'extraction des connaissances étant l'un des principaux objectifs de la bio-informatique [1].

Définition 2 : la bio-informatique est la science de l'informatique appliquée à la recherche biologique. De plus, la bio-informatique peut être décrite comme l'application de méthodes computationnelles pour la découverte de connaissances biologiques. Elle représente une symbiose de plusieurs domaines différents de la science, notamment, l'informatique, la biologie, les mathématiques et les statistiques [4].

Définition 3 : La bio-informatique est un domaine de recherche supra disciplinaire qui associe la biologie, la chimie, la physique et l'informatique en une science à socle large, importante pour approfondir notre compréhension des sciences de la vie [1].

Ainsi, c'est une discipline complémentaire aux approches classiques de la biologie [5] :

- In vivo (en français « en vie ») : tests au sein des organismes vivants.
- In situ or in utero « sur place » : tests dans les milieux naturels, souvent utilisées pour d'effectuer une expression impliquant un ou plusieurs organismes vivants. Elle fait référence au silicium (In silico).
- In vitro « en éprouvette » : tests dans des tubes.

In Vitro :
Analyse au niveau cellulaire

In Vivo :
Etudier les espèces modèles

In Silico :
Modélisation computationnelle

Figure I.2 : Les approches classiques d'expérimentation dans la biologie [5]

I.2.2 Historique de la Bio-informatique

Le tableau suivant présente une liste synthétisée [3][6] non exhaustive des principaux événements dans l'historique du développement de la bio-informatique.

Année	Événements majeurs
1944	Démonstration de l'ADN en tant que support de l'information génétique (Avery)
1946	ENIAC, premier ordinateur totalement électrique et Turing-complet
1947	Invention du transistor (Bell Labs)
1951	UNIVAC, premier ordinateur commercialisé
1953	Détermination de la première séquence protéique (insuline (Sanger, Thompson))
	Watson et Crick décrivent la structure en double hélice de l'ADN
1958	Première structure 3D de protéine (myoglobine (Kendrew))
	Réalisation du premier circuit intégré (Texas instruments)
1960	Énoncé du lien de causalité entre séquences et structures (Kendrew, Perutz)
1965	Premiers ordinateurs « de série » (IBM/360)
	Concepts de la phylogénie moléculaire proposés par Zuckerkandl et Pauling
1966	Premier modèle de structure 3D de macro-molécules sur ordinateur (Levinthal)
1967	Algorithme de construction d'arbres phylogénétiques (Fitch et Margoliash)
1970	Algorithme de recherche de similarité entre séquences (Needleman, Wunsch)
1971	Premier microprocesseur (Intel 4004)
1972	Clonage de fragments d'ADN dans un virus (Berg, Jackson, Symons)
1973	Découverte des enzymes de restriction qui coupe spécifiquement l'ADN
	Début du « Génie génétique » (Cohen, Boyer)
1974	Algorithme de prédiction des structures secondaires des protéines (Chou, Fasman)
1977	Mini-ordinateur DEC-VAX
	Premiers micro-ordinateurs (Apple, Commodore, Radioshack)
	Techniques de séquençage pour ADN (Sanger, Staden, Maxam, Gilbert)
	Premier génome complet à ADN séquencé : le bactériophage ϕX174 (Smith)
1980	EMBL : première banque de séquences nucléiques
1981	Séquence du génome mitochondrial humain (Anderson)
	Commercialisation du premier PC (IBM PC 5150)
	Algorithme d'alignement local de séquence (Smith-Waterman)
1982	GenBank (Los Alamos) contient 270 séquences pour 370 000 nucléotides
1984	Amplification de l'ADN : réaction de polymérisation en chaîne (PCR - Karry Mullis)
	Premier micro-ordinateur avec une interface graphique & souris (MacIntosh)
1985	Programme Fasta pour l'alignement local de séquences (Pearson-Lipman)
1987	Création du vecteur de clonage YAC (Yeast Artificial Chromosome) (Burke)
	Applied Biosystems : premier séquenceur automatique (ABI 370)
	Apparition de la technologie des puces à ADN (Kulesh)

1988	Lancement du projet international de séquençage du génome humain
	CLUSTAL, programme d'alignement multiple (Higgins, Sharp)
1990	Première identification d'un gène de maladie génétique par clonage positionnel et séquençage (premier essai de thérapie génique : neuro-fibromatose de type 1)
	Programme Blast de recherche de similarité entre séquences (alignement local) (Altschul)
1991	Expressed Sequences Tags (EST) : méthode rapide d'identification des gènes (Adams)
	GRAIL : programme de localisation de gènes dans le génome humain (Roberts)
1992	Séquençage complet du chromosome III de levure
1993	Création du European Bioinformatics Institute (EMBL - EBI)
1995	Premier organisme vivant séquencé : Haemophilus influenzae (Fleischman)
	Analyse du transcriptome : début des puces à ADN
1996	Premier organisme eucaryote séquencé : Saccharomyces Cerevisiae (Walsh, Barrell)
1997	Génome complet d'Escherichia coli (Blattner). 11 génomes bactériens disponibles
	Evolutions de BLAST (Altschul) : Gapped BLAST et PSI-BLAST
1998	Premier organisme pluricellulaire séquencé : *Caenorhabditis elegans*
2000	Séquençage du 1er génome de plante : Arabidopsis thaliana (Dennis, Surridge)
	Séquençage du génome d'une insecte : *Drosophila melanogaster (Adams)*
2001	Publication préliminaire du génome humain par Human Genome Project (Lander)
	Publication préliminaire du génome humain par *Celera Genomics (Venter)*
2002	Séquence préliminaire du génome de la souris (Waterson)
2004	Finalisation du génome humain (IHGSC)
	Projet ENCODE pour identifier tous les éléments fonctionnels du génome humain
2005 - 2008	Avènement des nouvelles technologies de séquençage à très haut débit (dites de 2^{éme} génération), ensuite de 3ème génération. Prise de conscience du phénomène "big data" (pas seulement en biologie) qui devient une discipline scientifique
2010	Création de la première bactérie synthétique, contrôler par un génome
2012	3708 génomes eucaryotes séquencés, plus de 14000 en projet (GenomesOnLine!) plus de 430.000.000.000 nucléotides !
2013	Nombre de génomes séquencés : 6880. Nombre de génomes en cours : 21058
2014	Plus de 776.000.000.000 nucléotides séquencés.
2015	Plus de 18.900 génomes eucaryotes et procaryotes séquencés, des milliers en projet. Plus de 1.363.000.000.000 nucléotides séquencés !
2021	≈941 milliards de Nucléotides & ≈191 millions Séquences d'acides aminés Plus de 22.600 génomes eucaryotes et procaryotes séquencés et des milliers en cours de séquençage (GenomesOnLine).
…	Voir (on line) le développement "spectaculaire" des banques de données : EMBL (Banque européenne créée en 1980), Genbank (Créée en 1982 en USA et diffusée par le National Center for Biotechnologie Information)

Table I.1 : Étapes clés du développement conjoint en bio-informatique [3][6]

I.2.3 Théories fondamentales de la biologie moléculaire

Le dogme central de la biologie moléculaire est un cadre permettant de comprendre le flux de l'information génétique au sein d'un système biologique. Ce dogme explique le processus par lequel l'information génétique est transférée de l'ADN à l'ARN, puis aux protéines, qui assurent les fonctions dans une cellule [7]. À travers la théorie fondamentale de la biologie moléculaire, les biologistes représentent un modèle schématique de la préservation et de l'utilisation de l'information génétique. En ce qui suit, on présente quelques définitions de base de la biologie moléculaire [8] :

- **Les cellules des Eucaryotes / procaryotes :** l'ensemble des organismes vivants peut être classé en trois grands groupes : les *eucaryotes,* les *eubactéries*, les *archaebactéries*. A l'intérieur de chacune de leurs cellules, les eucaryotes possèdent un noyau : petit sac entouré d'une membrane semi-perméable renfermant les chromosomes. L'Homme, ainsi que les animaux, les plantes et les champignons, sont des eucaryotes. Les eubactéries et les archaebactéries ne possèdent pas de vrai noyau, mais une structure beaucoup plus simple, non entourée d'une membrane. C'est de ce « proto-noyau » que vient le nom générique qui les désigne : procaryotes.
- **Exon / intron / gène mosaïque :** chez les eucaryotes, les gènes sont le plus souvent constitués de deux types de séquence nucléotidique : l'une est dite codante et l'autre non codante. Les parties codantes, appelées exons, portent l'information qui sera directement utilisée pour fabriquer les protéines. Entre les exons se trouvent les introns, non « lus » lors de la *traduction*. Du fait de cette disposition alternée exon/intron, on emploie l'expression *gène mosaïque*.
- **L'ADN (Acide Désoxyribo-Nucléique) :** l'ADN constitue le support biochimique de l'information génétique chez l'ensemble des êtres vivants, à l'exception de certains virus qui utilisent l'ARN. En tant que principal élément des chromosomes, l'ADN se présente généralement sous la forme de deux longs filaments (ou chaînes) enroulés l'un autour de l'autre, formant ainsi une structure en double hélice. Chaque chaîne est un polymère constitué de quatre nucléotides distincts, identifiés par la première lettre de leur base azotée : A (Adénine), C (Cytosine), G (Guanine) et T (Thymine).
- **L'ARN (Acide Ribo-Nucléique) :** dans les cellules, on identifie différents types d'ARN selon leur rôle. Les trois principaux types sont : les ARN messagers ARNm, les ARN de transfert ARNt et les ARN ribosomaux ARNr. L'ARN est un acide nucléique formé d'une unique chaîne de nucléotides, avec une structure similaire à celle de l'ADN. Toutefois, des différences chimiques entre ces deux types d'acides nucléiques confèrent à l'ARN des caractéristiques spécifiques. L'ARN est synthétisé via la transcription de l'ADN.
- **Protéine :** l'un des quatre matériaux de base de tout organisme, avec les glucides, les lipides et les acides nucléiques. Les protéines sont formées d'un enchaînement spécifique d'acides aminés (de quelques dizaines à plusieurs centaines). Le protéome est l'ensemble des protéines produites à partir du *génome* d'un organisme.

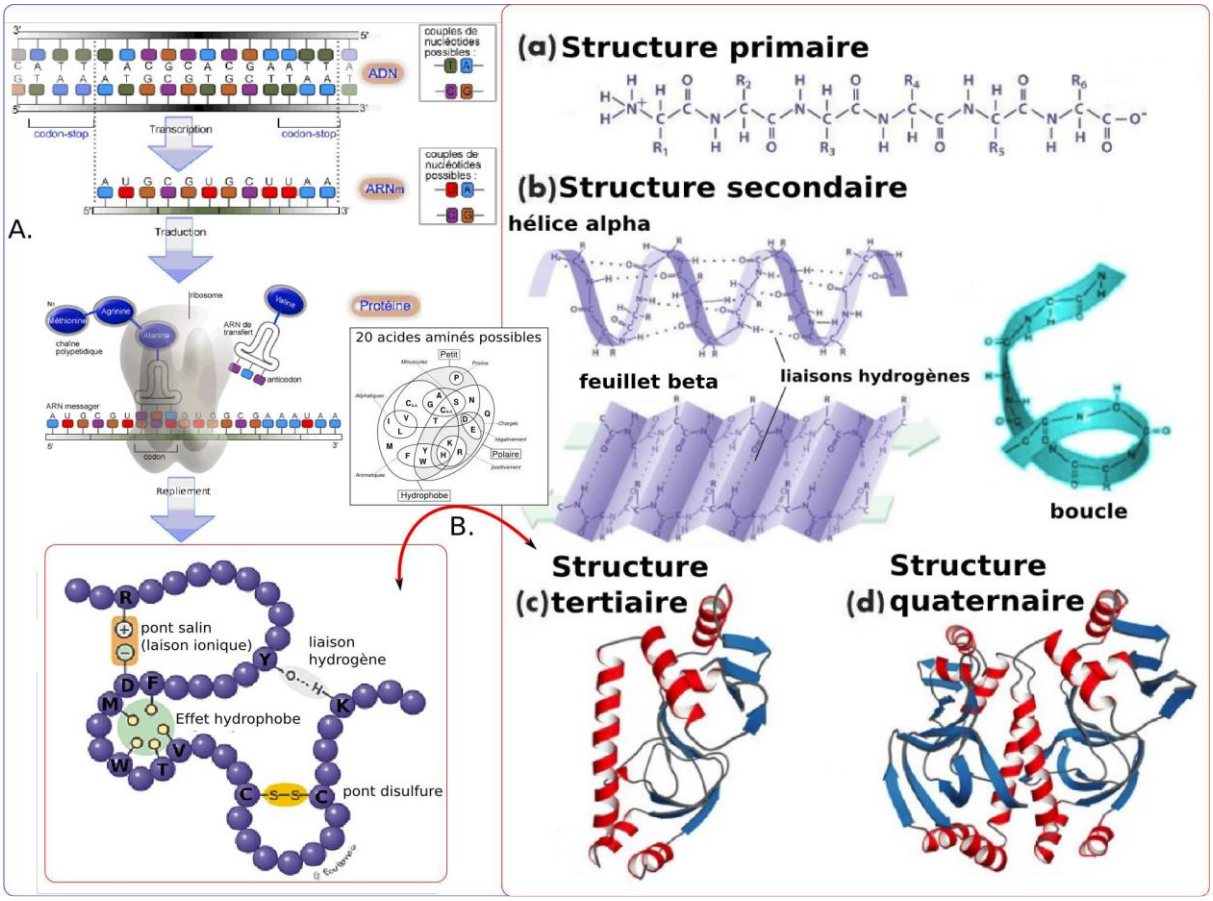


Figure I.3 : (A) - Schéma général de la synthèse des protéines : transcription de l'ADN en ARN messager (ARNm), puis traduction de l'ARNm en séquence d'acides aminés choisis parmi les 20 existants, classés ici selon leurs propriétés physico-chimiques (diagramme de Venn).
(B) - Du fait de la nature chimique des acides aminés et les interactions en résultant, la protéine se replie dans l'espace. Cette forme 3D repliée se divise en quatre niveaux de structure : (a) primaire, (b) secondaire, (c) tertiaire ou 3D et enfin (d) quaternaire [7].

Aussi, les éléments de base de la génétique moléculaire sont [8] :

- **Gène :** c'est un fragment d'ADN contenant les informations requises pour la synthèse d'une ou plusieurs protéines. Un gène inclut la séquence de nucléotides qui sera transcrite et ensuite traduite en acides aminés, ainsi que des séquences régulatrices qui contrôlent cette synthèse protéique selon les conditions de la cellule. La taille d'un gène peut aller de quelques centaines à plus d'un million de nucléotides.
- **Génome :** représente l'ensemble des informations génétiques d'un organisme. Une reproduction de ce génome se retrouve dans chacune de ses cellules. Il est hérité de génération en génération. De plus, le terme génome désigne également le support matériel de cette information génétique, à savoir la macromolécule d'ADN. Le génomique est l'étude des génomes (séquences d'ADN d'un organisme) et de ces fonctions.
- **Génomique fonctionnelle (« post-génomique ») :** étude la fonction des *gènes* par l'analyse de leur séquence et de leurs produits d'expression : les *ARNm* (*transcriptome*) et les *protéines* (*protéome*). Elle s'intéresse à leur mode de régulation, et à leurs interactions. L'analyse des protéines peut déterminer leur structure tridimensionnelle.

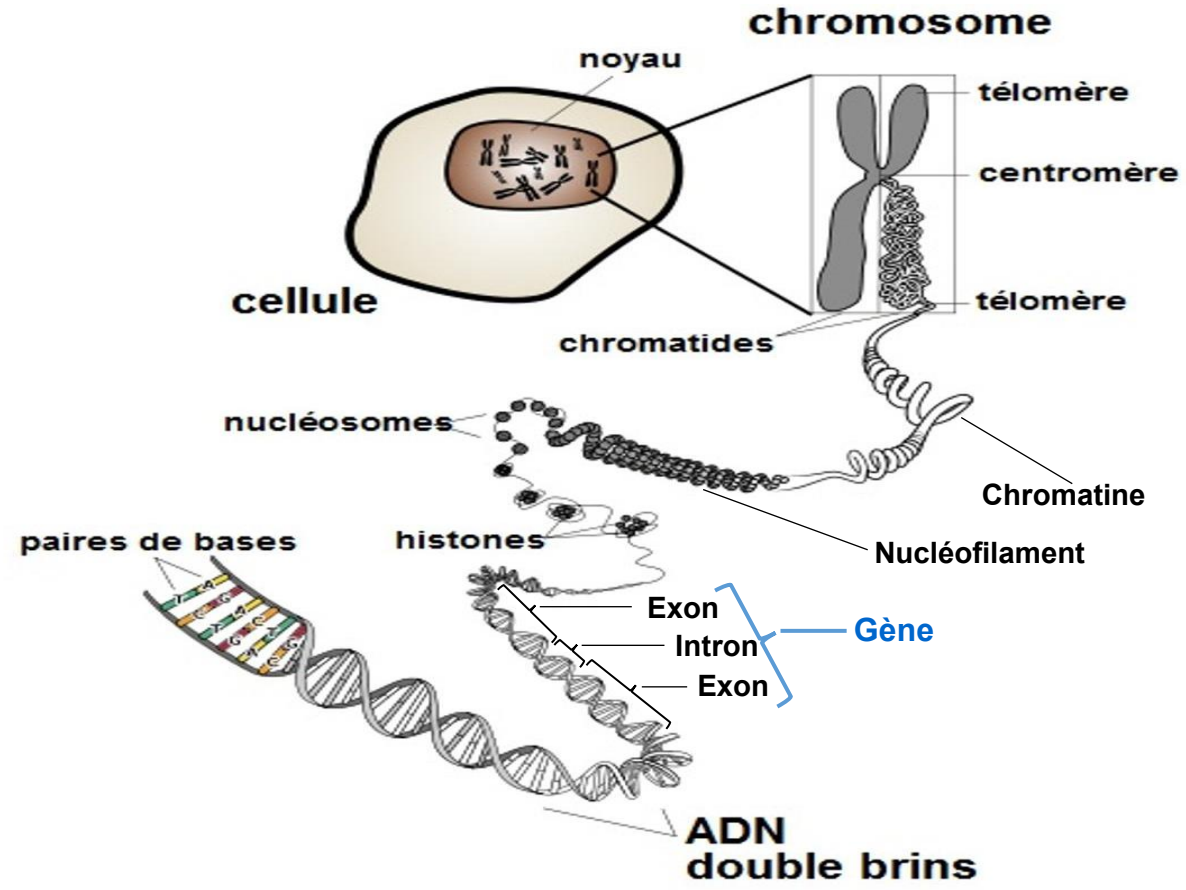


Figure I.4 : De l'ADN à la vie [65]

I.2.4 Les objectifs et domaines d'application de la bio-informatique

Les objectifs de la bio-informatique peuvent se résumer dans les points suivants [4] :

- Collecter et stocker des informations dans des bases de données :
  - Concevoir des outils de stockage adaptés à l'explosion de la quantité de données biologiques.
  - Fournir des outils divers facilitant la consultation et la visualisation
- Fournir des outils de comparaison de séquences protéiques et nucléotidiques :
  - Identifier une séquence en la comparant aux séquences d'une base de données.
  - Déterminer le degré de similitude entre deux séquences.
  - Repérer des motifs structuraux.
- Fournir des outils de traduction de séquences :
  - Simplifier les tâches de traduction, avec possibilité de proposer plusieurs traductions protéiques pour une même séquence.
  - Repérer les exons (fragments d'un ARN primaire qui se retrouvent dans l'ARN cytoplasmique après épissage) et les introns (fragments d'ARN primaire éliminés au cours de l'épissage).
- Acquérir une compréhension approfondie de la génétique des vivants pour :
  - Aider à comprendre les maladies complexes
  - Faciliter la conception de médicaments personnalisés.

Les principaux domaines d'application de la bio-informatique sont [9] :

- La bio-informatique des séquences : permet d'analyser les données dérivées de l'information génétique présente dans les trois types de séquences. Ce domaine se concentre particulièrement sur l'analyse et la comparaison multiples des séquences, l'identification de séquences à partir de données expérimentales, la détection des similarités entre les séquences, ainsi que la classification et la régression des séquences. De plus, elle vise à identifier les gènes ou les régions biologiquement significatives dans l'ADN ou les protéines, en s'appuyant sur les éléments de base (nucléotides, acides aminés).
- La bio-informatique structurale : concerne la reconstruction, la prévision ou l'analyse de structures à l'aide d'outils informatiques.
- La bio-informatique des réseaux : se concentre sur les interactions entre gènes, protéines, cellules et organismes.

Ainsi, la tâche majeure de la bio-informatique est de permettre d'identifier les fonctions d'un gène ou d'une protéine à partir de données existantes. Il existe différentes techniques aident à une meilleure compréhension de la fonction des gènes et des protéines, comme [4] :

- La construction évolutive d'arbre phylogénétique
- Détection de motifs dans les séquences
- Déterminer des structures 3D à partir de séquences
- Déduction de la régulation cellulaire
- Déterminer la fonction de protéine et les voies métaboliques
- Assembler les fragments d'ADN

I.3 Les banques de données biologiques

Le concept le plus important dans la bio-informatique appliquée est la collecte de données de séquence et son information biologique associée. Les projets de séquençage du génome génèrent quotidiennement une très grande quantité de données dans le monde entier. Pour utiliser ces données de façon appropriée, un système de dépôt structuré de ces données est nécessaire, mais les données devraient également être accessibles aux personnes intéressées. De ce fait, des bases de données en ligne ont vu le jour, elles sont accessibles gratuitement à tous, ce qui représente un avantage considérable. Différents chercheurs à travers le monde peuvent travailler sur le même ensemble de données et ainsi pouvoir comparer leurs résultats.

Ce concept a tellement pris de l'ampleur, que le journal *Nucleic Acids Research* consacre chaque année un numéro entier à toutes les bases de données biologiques disponibles, qui sont enregistrées sous forme de tableau avec les liens respectifs. En outre, pour un certain nombre de bases de données, des articles originaux décrivent leurs fonctions. Selon le type de données qu'elles contiennent, les bases de données biologiques peuvent être réparties en bases généralistes, ou bases spécialistes comme les bases d'images d'expression génétique [4].

I.3.1 Les banques de données généralistes

C'est un ensemble hétérogène d'informations mais le plus exhaustif possible, on cite [4] :

1. Banques de séquences nucléiques : il existe plusieurs banques généralistes de séquences nucléotidiques accessibles au public, on cite :
   - La banque EMBL (European Molecular Biology Laboratory), fondée en 1980 à Heidelberg, est gérée depuis 1994 par l'EBI (European Bioinformatics Institute).
   - GenBank : créé en 1979 au LANL (Los Alamos National Laboratory), maintenu depuis 1992 par le NCBI (National Center for Biotechnology Information).
   - La DDBJ (DNA Data Bank of Japan) a démarré ses activités en 1984 et a été créée et est toujours gérée par le NIG (National Institute of Genetics) à Mishima.
2. Banques de séquences protéiques, comme :
   - Swiss-prot : Fondée en 1986 par Amos Bairoch à Genève, cette banque est gérée en collaboration avec le SBI (Swiss Institute of Bioinformatics) et l'EBI.
   - PIR : La banque PIR (Protein Information Resource) a des origines anciennes, la première version ayant été lancée au milieu des années 60.

I.3.2 Banques de données spécialistes

Ce sont des bases de données centrées sur un thème précis : un type de molécule, une fonction biologique ou une maladie. Une des raisons de leur création réside dans la nécessité d'intégrer des informations spécifiques qui ne trouveraient pas leur place dans les collections généralistes. Parmi les banques spécialistes en biologie [4] :

- PDB (*Protein Data Bank)* : banque de données des structures tridimensionnelles des protéines et des acides nucléiques.
- KEGG (*Kyoto Encyclopedia of Genes and Genomes) : b*ase de données des voies métaboliques, fonctions biologiques et interactions moléculaires.
- OMIM (*Online Mendelian Inheritance in Man) : b*ase de données des maladies génétiques humaines et de leur hérédité.

I.3.3 Les séquences d'images d'expression génétique

Correspondent aux données qui décrivent quels gènes sont exprimés, où, quand et à quel niveau dans une cellule, un tissu ou un organisme. Parmi les bases des données d'expression génétique on trouve [4] [10-12] :

- Les séquences liées à l'expression génétique : comme les séquences d'ARN messager issues de gènes exprimés. De ce type, on peut citer les banques : GEO (Gene Expression Omnibus), ArrayExpress, ENA/GenBank (données RNA-seq)
- Les images d'expression génétique : qui sont des représentations visuelles montrant la localisation spatiale et temporelle de l'expression des gènes. Les types d'images ciblés sont : Microarrays (puces à ADN), RNA-seq visualisé (heatmaps), Hybridation in situ. Aussi, on peut citer les banques : Allen Brain Atlas (expression génique dans le cerveau), Emage Atlas (souris d'Edinburgh), Drosophila FlyBase, Zebra Fish Plateform.

Partie 2 : Les images d'expression génétique

I.4 L'imagerie in vivo d'expression génétique

Imagerie in vivo : comprend l'observation directe et en temps réel de ce qui se passe à l'intérieur d'un organisme vivant. Le procédé est principalement utilisé dans les laboratoires d'analyses et de recherche médicale ou biomédicale, mais est également utilisé dans les industries pharmaceutiques, alimentaires et chimiques, notamment pour étudier les effets de certains produits ou pour observer l'évolution de maladies ou de tissus. Les modèles d'expression génétique sont étudiés dans tous les principaux systèmes modèles animaux de manière systématique et les données résultant des études d'expression génique sont stockées dans une gamme de bases de données d'organismes modèles. Les principales bases de données sont FlyBase pour la drosophile, EMAGE pour la souris et ZFIN pour le poisson zèbre [10-12].

I.4.1 La banque *Drosofila FlyBase*

FlyBase (flybase.org) est l'une des bases qui possède le plus d'informations sur les interactions entre gènes de la Drosophile. Cette base a été fondée l'été 1992 par le National Institute of Health (USA) et le Medical Research Concil (UK) à partir du catalogue créé par Dan Lindsley et Georgianna Zimm : « The Genome of Drosophila melanogaster ». Une étude de la structure de la base de données a été faite. Celle-ci montre qu'elle est constituée d'entrées. Chaque entrée correspond à un gène donné. Chaque gène est décrit dans différents champs [12].

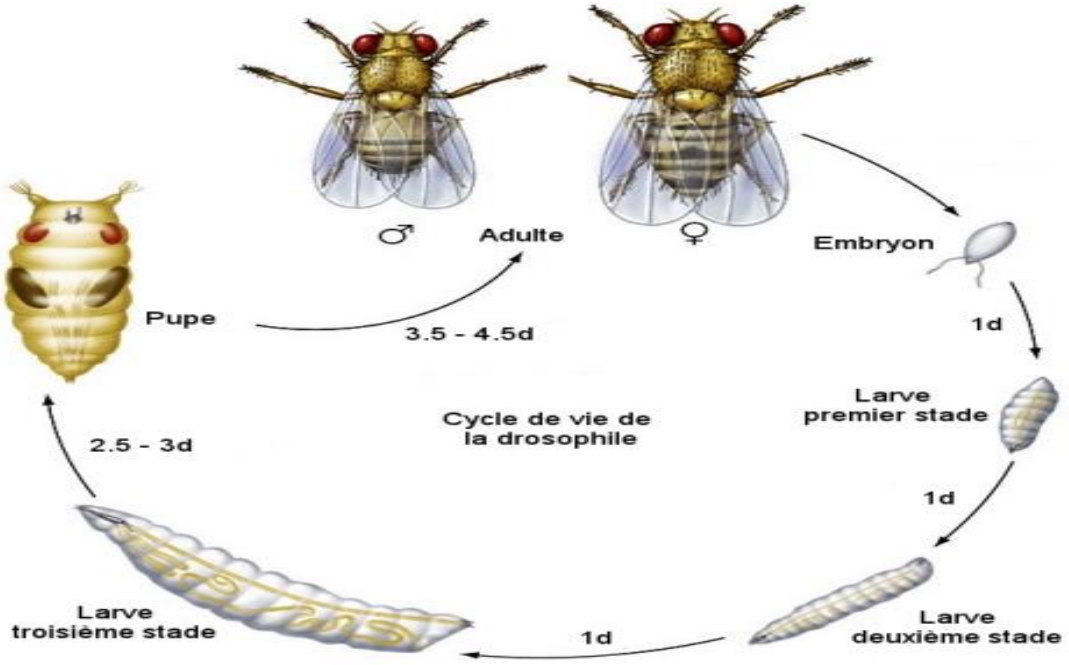


Figure I.5 : Cycles de vie de la Drosophile [66]

I.4.2 La banque *Zebra Fish*

Le poisson zèbre (Danio rerio) est un petit cyprinidé d'eau douce originaire du sous-continent indien. Le ZebraFish est l'un des organismes (espèces) modèles les plus célèbres pour les études en biologie des vertébrés et en génie génétique, car son embryon est transparent et donc les zones d'expression génétiques seront plus faciles à détecter et à cerner [11].

Figure I.6 : Image d'un poisson-zèbre [11]

L'Atlas de référence pour le poisson-zèbre est fourni dans la plateforme "ZebraFish Information Network" (ZFIN) (https://zfin.org/) qui fournis un large éventail de données génétiques et génomiques sur le poisson-zèbre, notamment les gènes, les allèles, les lignées transgéniques, l'expression et la fonction des gènes, les phénotypes mutants, l'orthologie, les modèles de maladies humaines, la nomenclature et les réactifs.

I.4.3 La banque *Edinburgh Mouse*

Les souris sont des espèces typiques importants de recherche pour modéliser la progression et le développement des maladies humaines en laboratoire, car les souris et les humains partagent des similitudes génétiques importantes. EMAGE (Edinburgh Mouse Atlas of Gene Expression) est un Atlas d'expression génique *in situ* durant les phases de développement de l'espèce modèle "Souris d'Edinburgh" [10]. La banque d'images EMAGE qu'on a utilisé est disponible sur le site https://www.emouseatlas.org/emage/ [13]. Cet Atlas d'images couvre les 23 jours (phases, nommées : Theiler Stages (TS)) de développement de l'espèce modèle "Edinburgh Mouse". La Figure I.7 représente les zones d'expression génique colorées suivant la force de cette expression : Expression Absente (Bleu), Faible (Mauve), modérée (Jaune), Forte (Rouge), et enfin détection Possible d'expression (Vert) [13].

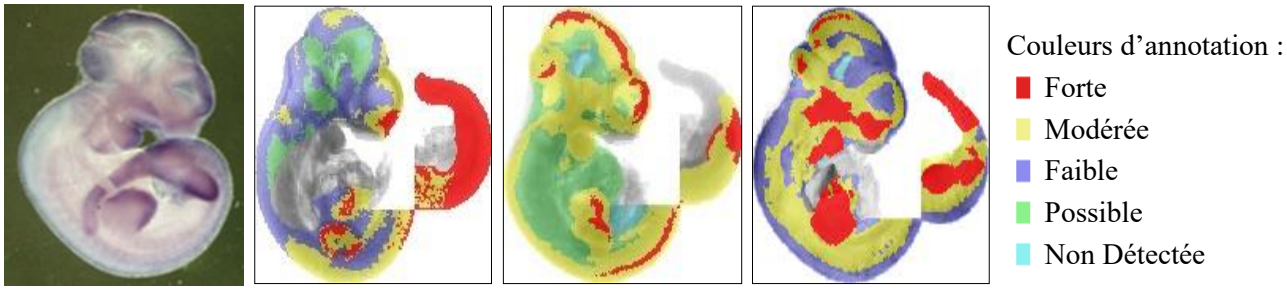


Figure I.7 : Exemple d'images d'Exp_Gen du géne nommé "ATF4" mappée sur l'image originale de embryon standard de la souris d'Edinburgh (image à gauche) durant les phases de development TS16, TS17, TS18 (Reference: EMAGE: 3052 in the online database: www.emouseatlas.org) [3]. Les zones d'Exp_Gen sont colorées en accordance avec le niveau de leur expression (voir la legende) [13]

I.5 Pourquoi choisir l'Atlas EMAGE

EMAGE est une base d'images d'expression génétique *in situ* de l'embryon de la souris d'Edinburgh, accompagnée d'une suite d'outils de consultation et d'analyse. La plateforme https://www.emouseatlas.org inclut des données d'hybridation in situ d'ARNm, d'immunohistochimie des protéines, de gènes rapporteurs transgéniques in situ et d'amplificateurs in situ. Aussi, cette plateforme traite les données provenant de la communauté scientifique, les décrit de manière standardisée, facilitant ainsi leur interrogation et leur échange. La description d'EMAGE comprend une composante textuelle, mais la particularité d'EMAGE réside dans son approche d'annotation spatiale des zones d'expression.

Dans notre étude on a utilisé la base d'images EMAGE pour étudier et modéliser caractéristiques des expressions génétiques. C'est un dataset gratuit d'images in situ à haute résolution, représentant les phases de développement de la souris, qui permet aux utilisateurs d'effectuer en ligne des requêtes multicritères sur l'expression des gènes. EMAGE est unique en fournissant à la fois des descriptions textuelles de l'expression génique et des cartes spatiales des modèles d'expression génique. Cette plateforme propose des outils avancés de : recherche de gènes/protéines (par : identifiant, stade de Theiler, type d'échantillon et type d'annotation, similarité de profils …), des données (images) spatiotemporelles sur d'expression génique in situ dans l'embryon de la souris, anatomie structurée, … etc [13].

Figure I.8 : La page accueil de la plateforme EMAGE [13]

I.6 Etat de l'art : travaux de recherche sur les données d'Exp_Gen

L'intérêt des données d'expression génétique, qui sont en rapide extension, est qu'elles aident à comprendre le rôle des gènes ainsi que leurs corrélations. On trouve dans la littérature beaucoup de travaux qui ont analysé, via des méthodes computationnelles, les données d'expression génétique dans leurs trois formes principales : Micro Array, mRNAseq et In Situ Hybridization (ISH). A cause de cette hétérogénéité, la compréhension et l'interprétation du contenu de ces données biologiques reste toujours un défi [14].

I.6.1 Micro Array data

La majorité des travaux sur ce type de données se concentrent en deux catégories : la recherche de distances et mesures de similarités adéquates pour assurer un bon clustering et découvrir les patterns cachés, et la sélection des features (gènes) les plus intéressants pour une classification appropriée des gènes.

I.6.1.1 Features selection

Le défi majeur lors de l'analyse des microarray data est le nombre limité d'échantillons (occurrences) par rapport au nombre élevé de features (gènes). Ainsi, l'objectif de la sélection des features est d'éliminer les gènes redondants ou non pertinents.

[15] propose une sélection de caractéristiques combinant des noyaux doubles RBF avec une analyse pondérée, pour extraire les features des données d'expression génique pour une classification des cancers. [16] proposent une sélection de caractéristiques basée sur les forêts aléatoires qui adopte l'idée de stratifier l'espace des features et combine des stratégies de séquence backward searching généralisé avec un sequence forward searching généralisé. Un score d'importance des variables basé sur les forêts aléatoires est utilisé pour classer les features.

Support Vector Machine (SVM) et ses variantes ont été intensivement utilisés pour la sélection de features et la classification. [17] se concentre sur la sélection des gènes pour la classification du cancer, en utilisant une optimisation binaire des essaims de particules (PSO) à comportement quantique pour sélectionner un sous-ensemble de gènes, SVM pour la classification du cancer et SVM avec leave-one-out cross pour la validation. [18] propose un modèle comprenant une sélection de sous-ensembles de features suivie d'une SVM multiclasses pour déterminer la classe fonctionnelle d'une séquence protéique nouvellement générée. Pour déterminer les features qui contribuent significativement à une classification fonctionnelle, des algorithmes de sélection comme Recursive Feature Elimination SVM (SVM-RFE) sont utilisés.

Aussi, pour sélectionner les gènes déterminants en relation avec le cancer du sein [19] propose un SVM-RFE avec optimisation des paramètres de la sélection des features via trois types d'algorithmes : grid search, les algorithmes génétiques, et enfin le PSO qui a démontré qu'il extrait les gènes les plus utiles.

Bien que le SVM-RFE ait prouvé son efficacité en feature selection, des améliorations ont été proposées. [20] avancent que l'inconvénient du SVM-RFE est son long temps

d'exécution. Ils proposent une implémentation plus efficace du SVM linéaire et une amélioration de la stratégie du RFE pour sélectionner des gènes informatifs. De plus, ils proposent un échantillonnage équilibré pour une classification plus crédible.

Aussi, [21] avancent que le SVM-RFE tente de trouver une classification binaire qui peut ne pas être biologiquement significative. Pour surpasser cette limitation, ils proposent sigFeature, un nouvel algorithme de features sélection, basé sur SVM et t-statistic pour découvrir les features différentiellement significatives afin d'améliorer la précision de la classification.

I.6.1.2 Clustering & mesures de similarité-distance

Pour obtenir des résultats utiles du clustering, on doit faire un choix « non biaisé » des paramètres comme le nombre de clusters ou leurs centres initiaux. De plus, on doit définir les mesures de similarité pour déterminer quelle distance utiliser.

[22] proposent un clustering des gènes basé non seulement sur la similarité entre leur force d'expression mais aussi suivant leur annotation fonctionnelle, obtenue depuis une Gene Ontology Database. Pour éviter le choix biaisé des hyper-paramètres [23] propose une mesure de validation des clusters basée sur le calcul de l'entropie d'une matrice de consensus pour déterminer le clustering le plus stable et estimer le nombre adéquat de clusters qui correspond à la réalité biologique des données analysées. [24] proposent une analyse du Fuzzy kernel Clustering pour identifier le nombre de clusters qui donne les segments les plus stables, avec un choix adéquat des distances et des centres initiaux des clusters.

I.6.2 mRNAseq

On trouve aussi des travaux de *features selection* pour la classification sur les données RNAseq, basés sur l'analyse différentielle des expressions génétiques pour étudier les changements d'expression dans les différentes conditions. Pour explorer l'utilisation des classificateurs multivariés pour la *features selection* des RNA-Seq, [25] utilise les mesures d'importance générées par des forêts aléatoires pour classer les gènes. [26] évaluent l'utilité des méthodes d'apprentissage supervisé sur les données RNA-seq pour une gamme diversifiée de classifications biologiques, avec l'hypothèse que les données d'expression au niveau de la transcription sont plus informatives pour la classification biologique que les données d'expression au niveau du gène.

Récemment, le Deep Learning a été largement utilisé pour analyser les profils d'expression génétique des données RNA-Seq. [27] proposent deux modèles de Deep Learning qui diffèrent dans la technique de features selection. Leurs résultats indiquent qu'une intégration des connaissances biologiques antérieures dans les modèles de Deep Learning permet d'obtenir de meilleurs résultats en termes de performances prédictives et de surpasser des modèles linéaires plus simples comme LASSO. De plus, parce que le dépistage précoce du cancer augmente les chances de guérison, [28] introduit une approche optimisée du Deep

Learning basée sur le Binary PSO avec les Arbres de Décision et Convolutional Neural Network pour classer différents types de cancer à partir des données RNA-seq tumorales.

Pour le clustering, [29] présente DESC, un algorithme de Deep Embedding qui segmente les données scRNA-seq en optimisant itérativement une fonction objective de clustering. Enfin, [30] présente DeepImpute, un algorithme d'imputation basé sur un réseau neuronal profond qui utilise des couches d'exclusion et des fonctions de perte pour apprendre des modèles dans les données RNA-seq unicellulaires.

I.6.3 Les images ISH

Etudier les images ISH fournit des patterns spatio-temporels détaillés d'expression génique des espèces modèles comme l'insecte Drosophila, la souris d'Edinburgh et le poisson zèbre. Dans la littérature, on peut distinguer trois principaux axes de recherche pour étudier les images ISH : la détection des gènes qui co-expriment, l'annotation des gènes et leur clustering pour extraire les patterns cachés.

En raison de la disponibilité en rapide croissance des données ISH, et des annotations inévitablement biaisées par les curateurs humains, il est nécessaire de développer des méthodes computationnelles pour analyser et interpréter les expressions génétiques.

Pour donner des classifications correctes au stade de développement des modèles d'expression génique embryonnaire de la drosophile, [31] ont extrait trois différentes features basées sur les modèles de mélange gaussien, l'analyse en composantes principales et les fonctions d'ondelettes, pour grouper les gènes qui semblent être co-régulés et partageant les mêmes modèles spatio-temporels d'expression génique.

[32] ont développé des méthodes de segmentation d'images pour identifier la similitude entre les modèles d'expression génique spatiale à partir de données ISH des embryons de la drosophile. Cette similitude est calculée via quatre métriques de scoring : erreur quadratique moyenne, distance d'ondelette de Haar, information mutuelle et information mutuelle spatiale.

[14] présente GINI, un système intelligent pour inférer des réseaux d'interaction génique à partir d'images ISH embryonnaires. Leur software s'appuie sur une représentation spatiale vectorielle (inspirée de la vision par ordinateur) du modèle spatial d'expression génique dans les images ISH. De plus, ils proposent un nouvel algorithme multi-instance-kernel qui apprend un modèle de sparse Markov network, dans lequel chaque gène (nœud) du réseau est représenté par un modèle spatial à valeurs vectorielles.

[33] proposent un modèle Tri-Relational Graph pour la classification des phases de développement de la drosophile ainsi que l'annotation des termes anatomiques des modèles d'expression génique. Enfin, pour construire des représentations 3D d'expression génique du cerveau, [34] a utilisé un réseau de neurones convolutionnel profond entraîné sur un grand ensemble d'images naturelles et appliqué pour extraire les features des images ISH de l'Atlas

Allen qui représentent le développement du cerveau de la souris, dans l'objectif d'annotation des patterns d'expression génique, ce qui démontre sa propriété de transfert d'apprentissage.

Les méthodes de clustering ont également été utilisées pour analyser les images ISH. [35] explorent la structure spatiale 3D des modèles d'expression génique du cerveau de la souris Allen à travers un ensemble de gènes consistants. Suite à une réduction de la dimensionnalité et du bruit, les voxels ont été segmentés suivant la similarité des modèles d'expression génique spatiale.

Pour faire face aux défis des données incomplètes, de la dimensionnalité élevée et de la complexité du transcriptome dans le cerveau, des méthodes de Deep Learning ont été utilisées. [36] proposent une solution au problème de la récupération des données. Les auteurs masquent une partie de chaque échantillon d'apprentissage afin que le Deep Neural Network puisse apprendre à récupérer les voxels manquants à partir du contexte. Ensuite, sur les données complétées, les auteurs ont utilisé des machines de Boltzmann restreintes pour inférer des cooccurrences depuis les voxels, fournissant des bases pour segmenter le cortex en sous-régions discrètes. Une représentation abstraite émerge des couches profondes du réseau de croyances correspondant aux structures anatomiques. [37] a entraîné un réseau de neurones convolutionnel pour dériver une mesure de similarité des images ISH de milliers de sections du tissu cérébral, et les incorporer dans un espace de features à dimension réduite, pour guider la conception de modèles d'enregistrement. Ensuite, ils ont compilé un atlas d'expression génique de la souris Allen.

I.7 Conclusion

Ce chapitre a présenté dans sa première partie les notions de base de la bio-informatique et les théories fondamentales de la biologie moléculaire, ainsi que leurs objectifs de recherche et domaines d'application. La deuxième partie était consacrée aux banques de données biologiques, et spécialement aux images d'expression génétique de quelques espèces modèles, pour se focaliser ensuite sur les images ISH d'expression génétique de la souris d'Edinburgh.

Dans le chapitre suivant, nous allons détailler le processus d'Extraction de Connaissances à partir des Données (ECD), et les catégories d'algorithmes de Machine Learning, pour se concentrer ensuite sur les méthodes d'extraction des règles d'association.

Chapitre 2

Chapitre II : Extraction des Connaissances à partir des Données (ECD) & Règles d'Association

II.1 Introduction

Le traitement et l'analyse des données volumineuses nécessite des méthodes hors du cadre classique basé sur des requêtes traditionnelles. Les systèmes informatiques classiques de gestion ne peuvent pas fournir les fonctionnalités nécessaires pour bien comprendre et exploiter ces données, afin de modéliser et d'extraire les connaissances cruciales qu'elles cachent.

Dans la première partie de ce chapitre nous présentons les différentes notions du processus d'Extraction des Connaissances à partir des Données (ECD), spécialement l'étape de fouille de données (Data Mining), ainsi que les algorithmes d'apprentissage automatique (Machine Learning) les plus utilisés. Dans la deuxième partie, nous expliquerons les concepts fondamentaux de l'extraction des règles d'association, pour nous concentrer ensuite sur les détails de l'algorithme de référence Apriori.

Partie 1 : Principes de base de l'ECD

II.2 Notions de base de l'ECD

II.2.1 Définition de l'ECD

Le terme ECD désignant l'extraction des connaissances à partir des données (*Knowledge Discovery in Databases : KDD*) est né des travaux croisés des chercheurs en statistique, en intelligence artificielle, et de visualisation des résultats. Ces travaux ont pour objectif l'automatisation de processus de découverte des connaissances pertinentes, nouvelles et utiles dans les grandes bases de données. On trouve dans la littérature plusieurs définitions de l'ECD, *Fayyad et al* [38] ont défini l'ECD comme *« le processus non trivial d'identification, à partir de données, de patterns valides, nouveaux, utiles et compréhensibles »*. D'après cette définition, on conclut que l'ECD est un processus qui comporte plusieurs étapes commençant par l'importation et la collection des données à partir des différentes sources pour enfin décrire des modèles ou patterns ou découvrir des relations entre les données. Le but est de rendre les informations plus lisibles et compréhensibles, pour découvrir des nouvelles connaissances qui sont utiles à la prise de décisions.

Une autre définition dans [39] annonce que : *« l'ECD vise à transformer des données (volumineuses, multiformes, stockées sous différents formats sur des supports pouvant être distribués) en connaissances. Ces connaissances peuvent s'exprimer sous forme d'un concept général qui enrichit le champ sémantique de l'usager par rapport à une question qui le préoccupe. Elles peuvent prendre la forme d'un rapport ou d'un graphique. Elles peuvent s'exprimer comme un modèle mathématique ou logique pour la prise de décision. Les modèles explicites quelle que soit leur forme, peuvent alimenter un système à base de connaissances ou un système expert ».* Cette définition est plus détaillée que la précédente, elle introduit un nouveau concept qui est celui de l'utilisation des connaissances extraites par un système de gestion des connaissances.

A partir de données volumineuses, le processus ECD vise l'extraction des connaissances, qui seront ensuite exprimés sous forme de concepts formels qui peuvent être des modèles mathématiques ou logiques, ou même une représentation graphique. Toutes ces formes de connaissances ont comme but commun, et le plus intéressant, l'aide dans la prise de décision.

II.2.2 Des données aux connaissances

Pour caractériser les différents types d'information ou de connaissance en fonction de leur complexité, de leur relation respective et de leur niveau de compréhension, on présente ses notions brièvement, aussi la figure II.1 schématise le lien entre ces différentes entités [40] :

- **Donnée** : élément de base dans une base de données (ou un Dataset en général), à savoir une mesure (numérique) ou une qualité (alphanumérique).
  - ✓ Exemple : 100
- **Information** : est une donnée interprétée qui représente un fait réel. C'est la donnée complétée par une description qui indique le contexte : de quelle mesure s'agit-il ?
  - ✓ Exemple : l'eau bout à 100° Celsius
- **Connaissance** : est une information assimilée, affinée et synthétisée se rapportant à un contexte spécifié. C'est l'information interprétable et exploitable, ayant un sens (pourquoi cette mesure a été prise ?).
  - ✓ Exemple : lorsque l'eau bout à 100°, les microbes seront éliminés
- **Compétence :** c'est une connaissance structurée pouvant être directement exploitée par les utilisateurs afin d'accomplir une certaine tâche ou action. Autrement dit, la compétence et la connaissance en action pour résoudre un problème réel.
  - ✓ Exemple : maîtriser les conditions d'ébullition de l'eau pour qu'il soit complétement stérilisé

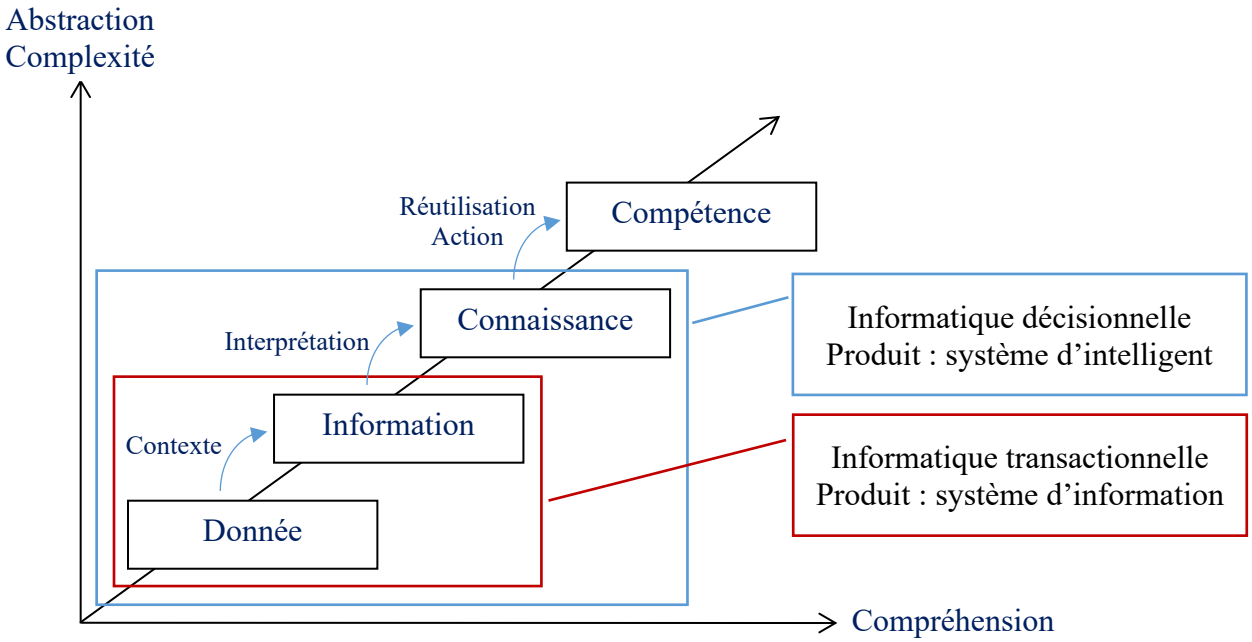


Figure II.1 : Des données aux connaissances [40]

II.2.3 Le processus d'ECD

Le processus d'ECD se déroule en quatre phases principales [38] [39] :

A. L'acquisition et organisation des données : consiste à organiser les données acquises afin de mieux les étudier. Cela via leur regroupement, intégration et fusion, structuration et mise en forme … etc. Les données peuvent être de différentes sources : bases de données, fichiers texte, fichiers XML, … et suite à leur acquisition on doit donc les organiser pour les stocker dans des récipients adéquats, généralement des entrepôts des données.

B. Le prétraitement et la transformation des données : c'est une étape très délicate et indispensable avant que les données soient disponibles pour l'analyse. En effet, les données issues de différentes sources présentent souvent diverses anomalies : doublons, valeurs manquantes (champs vides) ou incorrectes (bruit), … etc. Pour avoir des données de bonne qualité, diverses techniques de préparation et de transformation des données existent afin d'éliminer les incohérences. Ces techniques sont basées sur des opérations de : nettoyage et sélection des données, normalisation, réduction des dimensions, choix de format adéquat … etc. L'étape de prétraitement très intéressante, car tout simplement, la qualité et la pertinence des connaissances extraites dépend très étroitement de la qualité des données étudiées.

C. La fouille de données (Data Mining) : c'est l'étape cœur dans le processus d'ECD, c'est un ensemble de méthodes et de techniques ayant pour objectif d'analyser de grands volumes de données, souvent complexes et hétérogènes, afin d'y identifier : des *patterns*, des *relations*, des *tendances* ou des *règles significatives*. Les modèles ainsi obtenus permettent d'extraire des connaissances pertinentes, utiles et auparavant *inconnues ou implicites*, pouvant ensuite être exploités pour la *prise de décision*. Dans cette étape, il est crucial de choisir une méthode de Data Mining appropriée aux données étudiées et à l'objectif de l'analyse (type de connaissances à extraire).

D. Interprétation et évaluation des résultats : cette étape finale vise à soutenir la prise de décision en fournissant des modèles clairs pour les utilisateurs. Puisque ces derniers ont besoin d'une explication (simple et interprétable) du modèle obtenu, en préférence via une représentation visuelle. Ainsi, on aura besoin d'une part de métriques d'évaluation pour mesurer la performance des méthodes d'ECD utilisées pour avoir confiance lors de la réutilisation du modèle apprit ; de l'autre part on doit fournir une interprétation claire des résultats pour les rendre exploitables par l'utilisateur final.

Selon le schéma de la figure II.2, on remarque que le processus d'ECD est itératif, car il fait des retours dans n'importe quelle étape, pour des besoins de correction et de validation afin d'obtenir des connaissances de qualité. Ces retours sont déclenchés par l'utilisateur, suivant les spécificités de l'algorithme de Machine Learning utilisé et des données analysées.

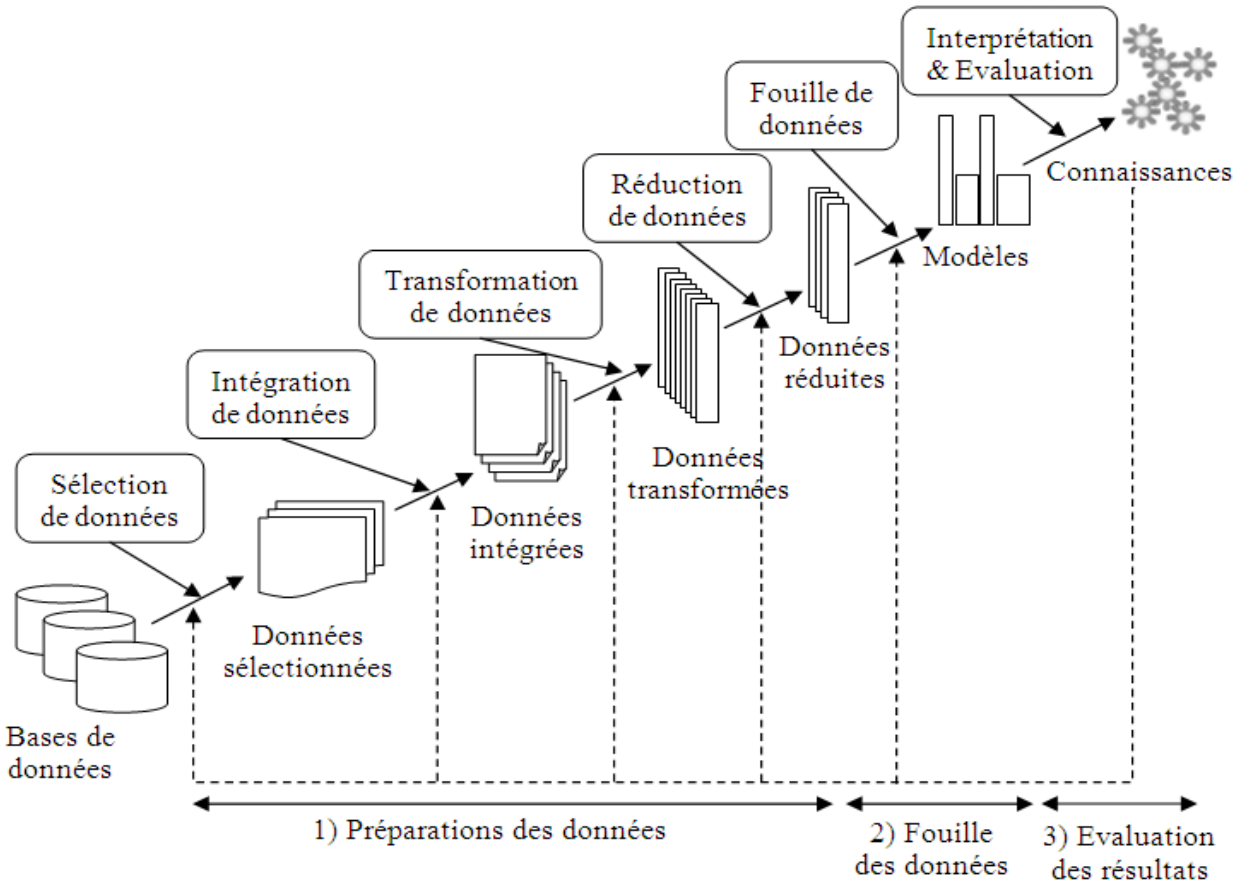


Figure II.2 : Processus d'ECD selon Fayyad et al [38]

II.3 Le Machine Learning

II.3.1 Qu'est-ce que le Machine Learning

Qu'est-ce qu'apprendre, comment apprend-on, et que cela signifie-t-il pour une machine ? La question de l'apprentissage fascine les spécialistes de l'informatique et des mathématiques tout autant que neurologues, pédagogues, philosophes ou artistes. Une définition qui s'applique à un programme informatique comme à un robot, un animal de compagnie ou un être humain est : « *L'apprentissage est une modification d'un comportement sur la base d'une expérience* ». Dans le cas d'un programme informatique, on parle d'apprentissage automatique, ou Machine Learning, quand ce programme a la capacité d'apprendre sans que cette modification ne soit explicitement programmée. On peut ainsi opposer un programme classique, qui utilise une procédure et les données qu'il reçoit en entrée pour produire en sortie des réponses, à un programme d'apprentissage automatique, qui utilise les données et les réponses afin de produire la procédure qui permet d'obtenir les secondes à partir des premières [41].

II.3.2 Principales tâches du Machine Learning

Sachant que chacune des méthodes de Machine Learning a ses propres métriques d'évaluation neutre de la pertinence de ses résultats, pour extraire un type précis de connaissances cruciales et utiles. Beaucoup de problèmes rencontrés peuvent être exprimés comme des tâches de [41] [42] :

A. Classification : c'est l'une des tâches les plus courantes en Machine Learning et une tâche essentielle pour l'humain. Elle consiste à examiner les caractéristiques d'un nouvel objet afin de l'assigner à une classe prédéfinie. Comme type d'apprentissage supervisé, les méthodes de classification se déroulent en deux phases : une phase d'entrainement qui s'appuie sur un jeu de données où la classe de chaque objet est déjà connue, dans le but de développer un modèle (formuler des règles pour attribuer des objets à des classes, en se basant sur des variables décrivant ces objets). Ensuite une phase de classification des nouveaux objets via le modèle appris.

B. Régression : est une modélisation permettant de faire des prévisions futures à partir de l'historique des données. On peut distinguer deux types de régression : linéaire et logistique. Le modèle de régression linéaire est basé sur des variables numériques expliquées et une variable explicative, il décrit le processus de détérioration à l'aide d'une équation linéaire. Tandis que la régression logistique est un modèle statistique qui permet d'analyser les relations entre un groupe de variables qualitatives X_i et une variable qualitative Y. Ce modèle est un linéaire généralisé qui utilise une fonction logistique comme lien. Cette régression permet aussi d'estimer la probabilité qu'un événement se produise (valeur de 1) ou non (valeur de 0) en optimisant les coefficients de régression. Les résultats obtenus se situent toujours entre 0 et 1.

C. Recherche d'associations : visent à identifier les variables qui sont liées dans une base de données. Un exemple classique est l'identification des produits (comme : pain et le lait, la tomate et les oignons) qui apparaissent ensemble sur un même ticket de caisse dans un supermarché. Cette analyse peut être réalisée pour déceler des opportunités de vente croisée et créer des assortiments de produits attractifs. Cette technique sera détaillée dans la deuxième partie de ce chapitre.

D. Segmentation : est une tâche d'apprentissage non supervisé car les données ne sont pas étiquetées et il n'existe pas de classes définies à priori. Son objectif est d'identifier des groupes cohérents d'individus partageant des caractéristiques similaires (découvrir les groupes (clusters) homogènes au sein d'une population). Le processus est relativement simple : il s'agit de maximiser la distance entre les classes tout en minimisant la distance (et donc maximiser la similarité) au sein des classes. Cette approche est plus complexe que l'apprentissage supervisé, car elle dispose de peu ou pas d'informations sur les données étudiées et sur les résultats attendus. Cette étape est souvent réalisée pour définir des groupes sur lesquels une classification pourra être appliquée.

E. Réduction des Dimensions : implique de passer d'un espace d'apprentissage de grande dimension à un espace de calcul plus compact qui engendre moins de *complexité algorithmique*. En d'autres termes, elle consiste à diminuer le nombre de variables ou de caractéristiques (features) utilisées pour entraîner le modèle d'IA. Les méthodes de réduction de la dimensionnalité se répartissent généralement en deux catégories : *la sélection de caractéristiques* (choisir les variables les plus pertinentes parmi celles du Dataset étudié) ; et *l'extraction de caractéristiques* (élaborer de nouvelles variables (caractéristiques) à partir des variables d'origine).

II.3.3 Le Machine Learning comme levier de l'IA

Les fondements de l'intelligence artificielle (IA) remontent au milieu du XX^{e} siècle, lorsque le mathématicien et logicien *Alan Turing* pose une question devenue emblématique : *« Une machine peut-elle penser ? »*. Cette interrogation marque un tournant majeur en introduisant une réflexion scientifique sur la possibilité de reproduire l'intelligence humaine à l'aide de machines. Le célèbre *test de Turing* constitue ainsi l'un des premiers cadres conceptuels pour évaluer le comportement intelligent d'un système artificiel. Parallèlement, les travaux de *Claude Shannon* sur la *théorie de l'information* jouent un rôle fondamental dans l'essor de l'IA. En établissant les bases mathématiques du codage, de la transmission et du traitement de l'information, Shannon fournit les outils nécessaires à la manipulation formelle des données, indispensables au développement des systèmes informatiques intelligents.

À partir de ces bases théoriques, l'intelligence artificielle se développe progressivement autour de plusieurs grandes approches. Parmi les premières, on trouve les *réseaux de neurones artificiels*, inspirés du fonctionnement du cerveau humain, les *systèmes symboliques*, reposant sur la manipulation de règles et de symboles logiques, ainsi que les *méthodes de raisonnement automatique*, visant à imiter la déduction humaine. L'IA s'impose alors comme un *champ scientifique multidisciplinaire*, dont l'objectif est de concevoir des systèmes capables de simuler des comportements intelligents tels que le raisonnement, la perception, la prise de décision ou encore la résolution de problèmes complexes.

Cependant, au cours des *années 1980,* l'IA connaît une période de *désillusion*, souvent qualifiée d'« hiver de l'IA ». Les *systèmes experts*, qui dominaient alors le paysage, se révèlent coûteux, rigides et difficiles à maintenir, et ne parviennent pas à répondre aux attentes industrielles et scientifiques. Cette crise entraîne une remise en question des approches classiques, notamment celles reposant sur la programmation exhaustive des connaissances.

C'est dans ce contexte que le *Data Mining* (qui est la partie fondamentale de l'ECD) a émergé dans les années 1990 comme une alternative prometteuse, cherchant à extraire automatiquement des connaissances à partir de grandes masses de données. Le *Machine Learning* peut ainsi être considéré comme l'héritier naturel du Data Mining : il marque un changement de paradigme majeur, passant de systèmes entièrement programmés à des modèles capables *d'apprendre, de s'adapter et de généraliser*. Ce renouveau méthodologique ouvre la voie à l'essor durable de l'intelligence artificielle moderne, tel que nous la connaissons aujourd'hui.

Dans ce contexte, le *Machine Learning (apprentissage automatique)* apparaît comme un sous-domaine central de l'IA. Contrairement aux approches traditionnelles basées sur des règles explicitement programmées, le Machine Learning vise à permettre aux machines *d'apprendre automatiquement à partir des données*, en identifiant des régularités et en construisant des modèles descriptifs ou prédictifs [42].

II.3.4 Algorithmes de Machine Learning

Suivant l'objectif de l'apprentissage, les techniques de l'IA peuvent être répartis en deux axes : les algorithmes de Machine Learning qui ont pour objectif d'identifier des structures ou des modèles mathématiques latents (hidden patterns) ; les algorithmes d'optimisation visent à déterminer, de façon exacte ou approchée, des solutions optimales à un problème, en se basant sur une fonction objective qui cherche des valeurs maximales (gains) ou minimales (couts). La figure II.3 présente un récapitulatif non exhaustif des principaux catégories et algorithmes de Machine Learning, en classant l'algorithme Apriori utilisé dans la sous-catégorie adéquate [42].

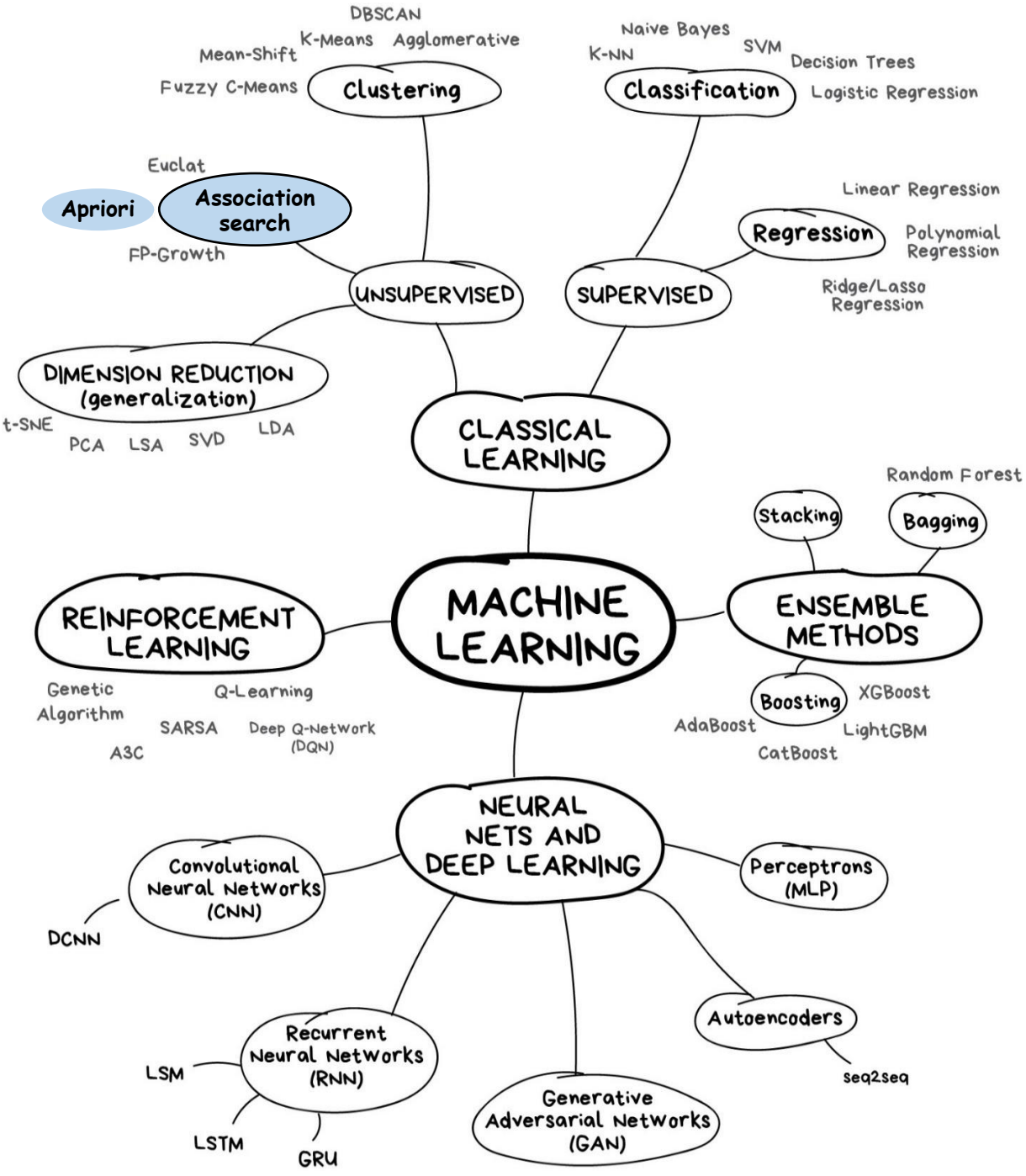


Figure II.3 : Récapitulatif graphique des différents algorithmes de Machine Learning [42]

Partie 2 : Autour et alentour des Règles d'Association

II.4 Les méthodes de recherche d'associations

II.4.1 Présentation générale

La recherche d'association et parmi les méthodes de recherche des implications (relation) entre les attributs d'une base de données, elle constitue un moyen pratique et efficace de découvrir et de représenter certaines dépendances et relations significatives entre attributs dans une base de données. L'étymologie des RA, aussi connues sous le nom de l'analyse du panier de la ménagère (*Market Basket Analysis*), vient des travaux de [43] qui ont été réalisés pour l'analyse les transactions d'achat. Chaque panier représente les articles achetés par un client dans une seule transaction d'achat. La méthode consiste à rechercher quels groupes de produits qui sont fréquemment achetés ensemble, elle fournit des informations sous forme de règles, comme par exemple : « Si un client acheté du lait, il achète du gâteau (70%) ». Cette règle signifie que 70% des clients qui achètent du lait ont également tendance à acheter du gâteau, cette découverte est très intéressante et peut être utilisée dans le but de prédire le comportement des clients (prédiction), comme elle peut aider un gestionnaire du supermarché à ranger ses étagères afin que les produits qui sont souvent achetés ensemble se trouvent à proximité. Il peut aussi décider quelles sont les articles à mettre en promotion (les articles rarement achetés).

II.4.2 Développement historique de la recherche d'associations

L'émergence de la fouille de données, une étape qui est située au cœur du processus de découverte des connaissances à partir des données (ECD) [38] dont le but est de découvrir des modèles cachés dans les Datasets, afin de les utiliser pour l'aide à la prise des décisions, parmi lesquelles l'analyse des liens. Cette dernière est accomplie par plusieurs techniques, y compris les RA, devenue l'une des plus importantes applications dans le domaine du *Machine Learning*. L'étude des associations entre attributs booléens est ancienne. L'une des premières méthodes de la recherche des RA est la méthode *GUHA* (General Unary Hypotheses Automaton) initiée par Hajek et al. en 1966 [44], où apparaissent les notions de support et de confiance. En 1986, *Guigues et Duquenne* [45] s'intéressaient plutôt à une représentation ordonnée de concepts avec les implications informatives. En 1993, *Agrawal et al* [43] ont développé ce qui va être la méthode de base pour d'extraction des RA dans les grandes bases de données. Des variantes de leur algorithme seront détaillées dans la section II.6. Aussi, un bon résumé sur l'extraction des RA est proposé dans [46], qui présente un aperçu des différentes problématiques de recherche liées à l'identification des associations positives (entre les items corrélés fréquemment), et les associations négatives (entre les items négativement corrélés, qui s'annulent mutuellement) qui sont largement négligés mais qui peuvent révéler des connaissances très utiles.

II.5 Concepts fondamentaux des règles d'association

II.5.1 Notions de base

L'extraction des RA se base sur plusieurs concepts fondamentaux, notamment [47] :

II.5.1.1 Item

Un item désigne tout article, objet ou attribut appartenant à un ensemble « I » fini d'éléments différents tel que I = {i_1, i_2, i_3, …, i_d}, où d est le nombre d'items.

II.5.1.2 Itemset

Un itemset X noté {$i_1 i_2 \ldots i_m$} désigne un ensemble de m items où i_j est un item de I. Par exemple {$i_2 i_6 i_9$} est un itemset compose de 3 items.

II.5.1.3 K-Itemset

Un k-itemset est une famille d'itemsets, où chaque itemset contient k items. Par exemple, l'itemset {$i_2 i_6 i_9$} est un élément de la famille 3-Itemset.

II.5.1.4 Support d'un itemset

Soit T = t_1, t_2, …, t_n une base de données de n transactions, où chaque transaction t_i, (i=1,2, …, n) est représentée par un ensemble d'items de I (c'est-à-dire, $t_i \subseteq I$), le support d'un itemset X noté (Supp(X)) est le ratio de toutes les transactions contenant l'itemset X divisé par le total de toutes les transactions.

$$Supp(X) = \frac{|t_i \in T / X \subseteq t_i|}{|T|} \qquad \text{(II.1)}$$

Où | | désigne la cardinalité d'un ensemble.

II.5.1.5 Règle d'association

Une règle d'association est une relation d'implication sous la forme de : X → Y, elle exprime le fait de si X alors Y, où X, Y ⊆ I. X est la condition (ou antécédent) et Y est la conclusion (ou résultat) de cette règle. X et Y n'ont pas un item commun (i.e. X∩Y = ∅).

II.5.1.6 Confiance d'une règle d'association

La confiance de la règle d'association X → Y est le ratio entre le nombre de transactions qui contiennent simultanément les deux itemsets X et Y divisé par le nombre de transactions contenant seulement l'itemset X.

$$Conf(X \rightarrow Y) = \frac{|t_i \in T \,/\, (X \subseteq t_i) \cap (Y \subseteq t_i)|}{|t_i \in T \,/\, X \subseteq t_i|} = \frac{Supp(X \cup Y)}{Supp(X)} \qquad \text{(II.2)}$$

II.5.1.7 Itemset fréquent

Un itemset est dit fréquent si et seulement si son support est supérieur ou égal à un seuil minimal fixé préalablement par l'utilisateur [43]. Pour un dataset contenant d items, le nombre des k-itemsets possibles (en faisant les combinaisons entre items) est présenté dans Tab II.1.

	d items	k=2	k=3	k=4
K-Itemsets possibles (C_d^k)	100	4 950	161 700	3 921 225
	10 000	5×10^7	1.7×10^{11}	4.2×10^{14}

Table II.1 : La complexité algorithmique élevée de la recherche des itemsets fréquents

II.5.2 Les étapes de l'extraction de règles d'association

Le processus de découverte des RA utiles se fait en quatre phases principales [48] :

- **Prétraitement des données :** c'est la préparation des données pour faciliter l'exploration des RA, via la sélection des attributs (items) les plus intéressants, cela permet de réduire la taille des données traitées. Après, on converti ces données en format binaire.
- **Extraction des itemsets fréquents :** c'est l'étape la plus couteuse en temps d'exécution et en espace mémoire, car le nombre d'itemsets possibles est de complexité exponentielle vis à vis le nombre d'items manipulés (voir Tab II.1). On doit donc réduire l'espace de recherche et sélectionner seulement les itemsets fréquents, ayant un support ≥ Min_Sup.
- **Génération des règles d'association :** c'est la découverte des RA les plus pertinentes à partir de l'ensemble d'itemsets fréquents qu'on a déjà extrait. Ces règles doivent donc aussi avoir une confiance supérieure à un seuil Min_Conf fixé par l'utilisateur.
- **Visualisation et interprétation des règles extraites :** en les présentant à l'utilisateur convenablement, pour qu'il puisse confirmer leur intérêt et leur pertinence. Aussi, une bonne interprétation des règles extraites engendrera une meilleure prise de décision.

La figure suivante illustre les différentes étapes d'extraction de règles d'association :

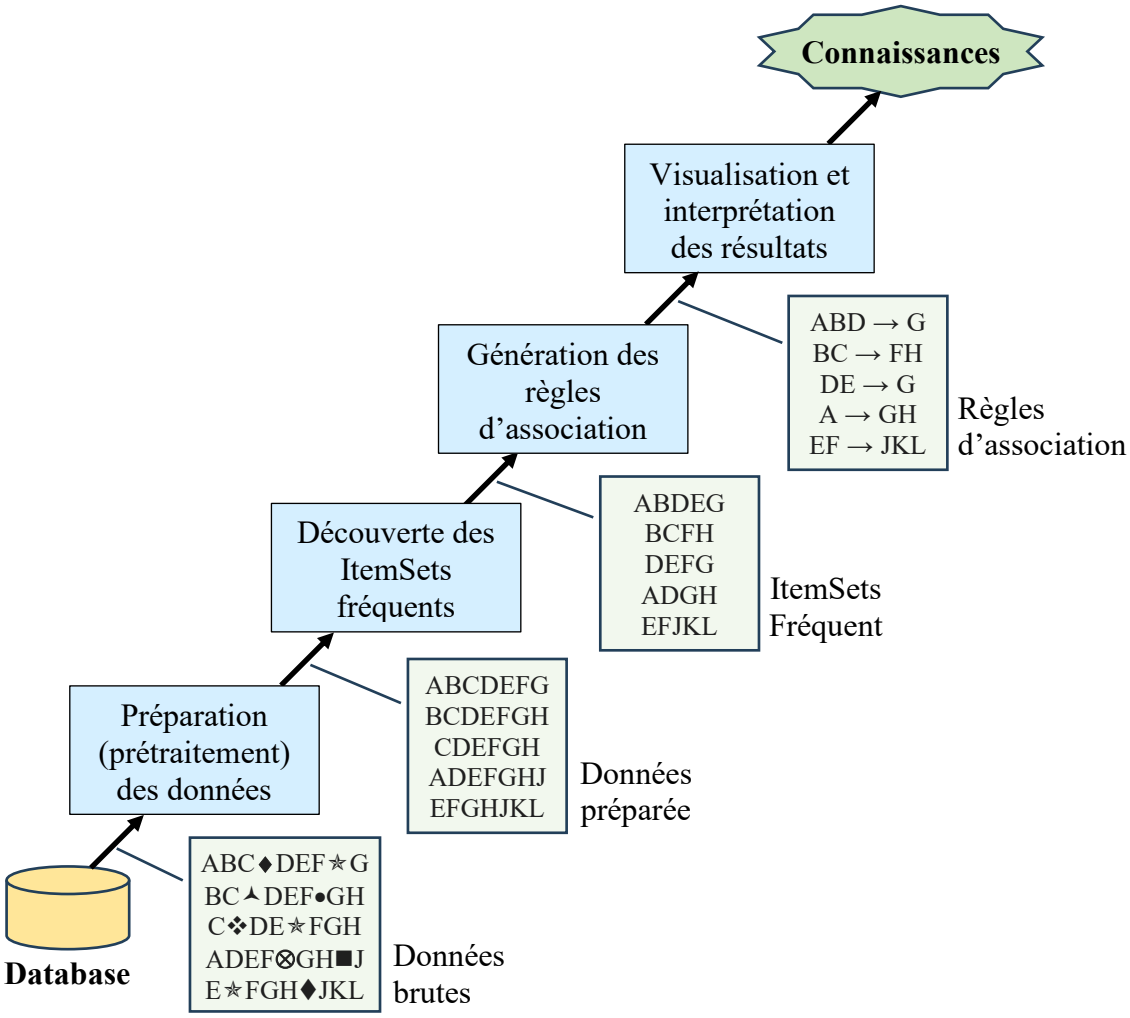


Figure II.4 : Les étapes d'extraction des règles d'association [48]

II.5.3 Espace de recherche ou le treillis des itemsets

Dans la recherche des RA, les bases de données sont vues comme une matrice binaire, chaque ligne est une transaction et les colonnes sont appelées items. Dans l'exemple du panier d'achats, une ligne est un passage à la caisse, une colonne est un article. La table Tab. II.2 contient un ensemble de transactions des items A, B, C, D. Le treillis des itemsets contient « $2^{|items|}$ itemsets », on associe à chaque itemset un indice pour indiquer son support. Le treillis comporte « N = |items| +1 » niveaux (voir Fig. II.5). Plus le nombre d'items est grand, plus l'espace de recherche est large, plus l'algorithme qui génère les itemsets fréquents est défaillant. Supposons qu'on cherche les itemsets fréquents à partir de Tab. II.2, avec un Min_Sup = 2. La ligne qui traverse le treillis est une frontière au-dessus de laquelle sont situées les itemsets non fréquents, et au-dessous les itemsets fréquents. Les algorithmes d'extraction des RA essaient d'avoisiner cette frontière pour élaguer le mieux possible l'espace de recherche inutile [47].

Id-Trans	Item A	Item B	Item C	Item D
01	1	1	0	0
02	1	1	1	0
03	1	0	1	1
04	1	1	0	1
05	0	1	0	1
06	0	0	1	1

Table II.2 : Table de transactions

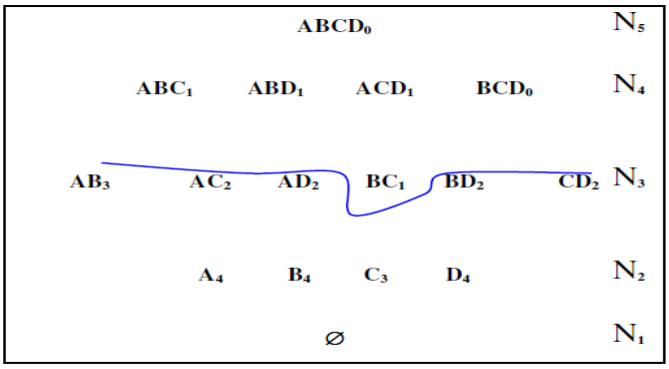


Figure II.5 : Treillis des items

Il y a deux façons de parcourir l'espace de recherche, en *largeur* ou en *profondeur*. Un algorithme de parcours en profondeur [47] considèrera d'abord les itemsets ayant les mêmes préfixes (par exemple {a}, {a, b}, {a, b, c, d}) ; alors qu'un algorithme de parcours en largeur (aussi appelé « horizontal » ou « par niveaux ») considèrera d'abord les itemsets de même taille. Ces deux ordres de parcours sont illustrés par la figure suivante (Fig. II.6). Dans le 1^{ier} cas, les supports de tous les itemsets de niveau (k-1) du treillis doivent être calculés avant de générer les itemsets d'un niveau k. Le 2^{éme} parcours consiste à construire un arbre (une projection de la base de données) de manière récursive tout en scannant la base de données.

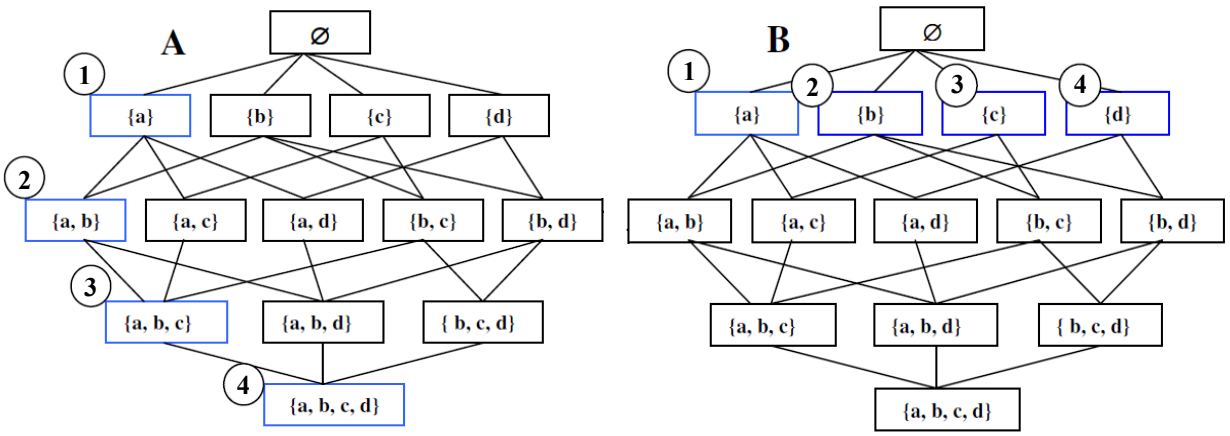


Figure II.6 : Algorithme de parcours des itemsets en profondeur (A) et en largeur (B) [47]

II.5.4 Mesures d'évaluation et de validation

La complexité d'extraction des RA est en fonction du nombre d'itemset fréquents générés. Pour *N* items fréquents on peut avoir (2^N-1) *Itemset fréquents*, ce qui conduit à produire un très grand nombre de règles, on aura donc le problème de choix des meilleures règles. Parfois ce processus génère des RA dites *triviales* ou *inutiles*. La première catégorie (triviales) représente les règles évidentes, c'est-à-dire celles qui sont déjà connu et qui n'apportent pas d'information de plus. Par exemple la règle : SI *achat d'une imprimante* ALORS *achat d'encre*. La seconde catégorie (RA inutiles) représente les règles difficilement interprétables.

On peut trouver dans la littérature plusieurs critères d'évaluations des règles d'association, qui sont divisées en deux catégories [49], la première dite *subjective* ou l'expert du domaine sait quels attributs il souhaite avoir dans les RA, la deuxième dite *objective* et consiste à étudier le nombre d'exemples et de contre-exemples. Les mesures de qualité objectives les plus utilisées dans les algorithmes d'extraction des RA sont : le support (appelé aussi mesure d'*utilité*), la confiance (appelée mesure de *précision*). En général, il faut que la règle obtenue ne soit pas trop générale ou évidente, car dans ce cas, il n'y a rien de nouveau. Elle ne doit pas non plus être trop spécifique, provenant seulement de données aberrantes.

Tous ces contraintes ont conduit les chercheurs à proposer plusieurs critères pour mesurer la qualité des règles, visant à améliorer les performances du processus d'extraction des RA. *Vaillant et al* [50] ont implémenté la plate-forme *HERBS,* ils ont utilisé vingt mesures de qualité, basées sur les quatre cardinalités de la RA ($A \rightarrow B$), ces cardinalités sont les fréquences d'apparitions de l'exemple et de contre-exemple de la RA, et on peut les exprimer par des probabilités de survenu des évènements (tables de contingence Tab. II.3.a et Tab. II.3.b). Les fréquences sont schématisées par la figure Fig. II.7.

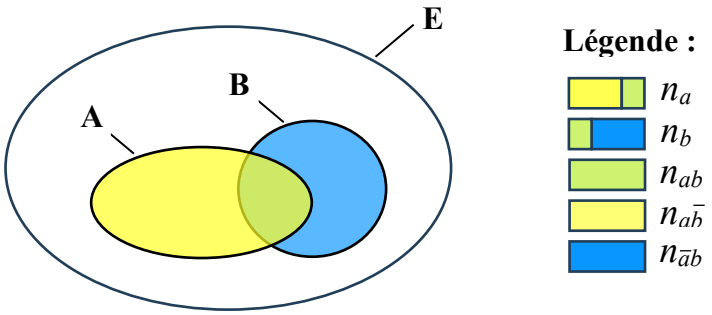


Fig.II.7 : Les fréquences d'apparition des itemsets A et B [50]

A \ B	0	1	Total
0	$n_{\bar{a}\bar{b}}$	$n_{\bar{a}b}$	$n_{\bar{a}}$
1	$n_{a\bar{b}}$	n_{ab}	n_a
Total	$n_{\bar{b}}$	n_b	n

(a)

A \ B	0	1	Total
0	$p_{\bar{a}\bar{b}}$	$p_{\bar{a}b}$	$p_{\bar{a}}$
1	$p_{a\bar{b}}$	p_{ab}	p_a
Total	$p_{\bar{b}}$	p_b	1

(b)

Table II.3 : Tables des contingences [50]

Voici une brève présentation de quelques mesures d'intérêts des RA :

A. Confiance centrée : calculée par formule : *confiance_centrée*(A→*B*) = P(B/A)-P(B)) [49], cette mesure permet donc de relativiser la confiance d'une règle (*A* →*B*) par rapport à la taille de sa partie droite B.

B. Rappel : aussi connue sous le nom de mesure de *sensibilité,* le *rappel* est calculé par : *rappel*(*A*→*B*) = *P*(*A*/*B*) [51]. Elle permet d'évaluer la proportion d'objets vérifiant la prémisse de la règle parmi ceux vérifiant la conclusion.

C. Lift : c'est le rapport entre l'indice brute P(AB) d'une règle et l'indice d'indépendance P(A)XP(B) entre les parties de cette règle. Calculé par : $\boldsymbol{Lift(A \rightarrow B)} = \frac{Conf(A \Rightarrow B)}{Supp(B)} = \frac{P(AB)}{P(A)P(B)}$

Le lift permet donc d'apprécier, pour une règle (A→B), sa « distance » à l'indépendance. Cette mesure est symétrique, elle ne distingue pas les règles (A→B) et (B→A) [52].

D. Pearl : calculée par : *Pearl*(A→B) = P(A) X |P(B/A) - P(B)|) , cette mesure permet d'évaluer l'intérêt d'une règle A→B par rapport à l'hypothèse d'indépendance entre la prémisse et la conclusion de la règle [49]. La mesure Pearl permet de vérifier que la règle apporte vraiment une connaissance nouvelle sur les données. En effet, si la mesure de Pearl de la règle est proche de 0, cela nous indique que les Items (A et B) sont liés par une relation intéressante.

E. Piatetsky-Shapiro : calculé formellement par : *PS*(A→B) = n X P(A) X (P(B/A) - P(B)). Cet indice est très proche de l'indice de *Pearl*, puisqu'il peut être exprimé de la manière suivante : PS(A→B) = n X *Pearl*(A→B). L'indice de Piatetsky-Shapiro [53] évalue l'intérêt d'une règle par rapport à son écart à l'indépendance.

Le tableau suivant présente les principales mesures de l'intérêt des RA :

Nom	Formule
Confiance	$P(B/A)$
Confiance centrée	$P(B/A) - P(B)$
Pearl	$P(A)\,\lvert P(B/A) - P(B)\rvert$
Piatetsky-Shapiro	$nP(A)\,(P(B/A) - P(B))$
Loevinger :	$\frac{P(B/A)-P(B)}{P(\overline{B})}$
Zhang :	$\frac{P(AB)-P(A)P(B)}{Max\{P(AB)P(\overline{B});P(B)(P(A)-P(AB)\}}$
Corrélation	$\frac{P(AB)-P(A)P(B)}{\sqrt{P(A)P(\overline{A})P(B)P(\overline{B})}}$
Indice d'implicat.	$\sqrt{n}\frac{P(A\overline{B})-P(A)P(\overline{B})}{\sqrt{P(A)P(\overline{B})}}$
Lift	$\frac{P(AB)}{P(A)P(B)}$
Surprise	$\frac{P(AB)-P(A\overline{B})}{P(B)}$
Conviction	$\frac{P(A)P(\overline{B})}{P(A\overline{B})}$
Intensité d'implicat.	$P\left[Poisson\left(nP\left(A\right)P\left(\overline{B}\right)\right) \geq nP(A\overline{B})\right]$
Sebag-Schoenauer	$\frac{P(AB)}{P(A\overline{B})}$
Multiplicateur de cote	$\frac{P(AB)P(\overline{B})}{P(A\overline{B})P(B)}$
J-mesure	$P(AB)\log\frac{P(AB)}{P(A)P(B)} + P(A\overline{B})\log\frac{P(A\overline{B})}{P(A)P(\overline{B})}$

Table II.4 : Présentation des principales mesures d'intérêt des RA [49]

II.6 Divers algorithmes d'induction des règles d'association

L'extraction des RA est l'un des algorithmes les plus utilisés pour la découverte des connaissances dans une source de données (ou l'ECD). Le but de cette technique est de découvrir les relations significatives entre les attributs caractérisant les objets. Pour cela, plusieurs algorithmes traitant le problème de recherche de règles d'association ont été proposés dans la littérature tel que l'algorithme AIS [43], Apriori [54], l'algorithme Partition [55], l'algorithme Eclat [56], l'algorithme FP-growth [57], Algorithme C5.0 [58], … etc. Chacun de ces algorithmes a ses spécificités et ces avantages-inconvénients, mais l'algorithme Apriori proposé par *Agrawal et al* [54] restera la référence dans ce domaine et même la base de quelques autres algorithmes. En ce qui suit on proposera une brève définition de ces algorithmes.

II.6.1 Algorithme C5.0

Son principe est de transformer un arbre de décision à un ensemble de règles explicites de la forme [*Si ... Alors ...*]. Donc à partir de la structure hiérarchique de l'arbre on va déduire des règles d'association décrivant les relations entre les variables et les classes de décision. Cette forme de modélisation est préférée car les arbres de décision sont lisibles et facilement explicables, ce qui permet une meilleure compréhension du processus de décision et facilite l'analyse et la validation des résultats obtenus [58].

II.6.2 Le GRI (Generalised Rule Induction)

C'est outil d'induction engendre un ensemble de règles indépendantes, contrairement aux ensembles des règles générés par C5.0 qu'ils sont disjonctifs (du fait qu'ils s'appuient sur un arbre de décision). Chaque règle comporte une partie « conditions » portant sur les champs disponibles et une partie « conclusion » portant sur des champs choisis au préalable l'utilisateur aura. De plus, chaque règle se voit affecter un taux de couverture et une précision [58].

II.6.3 Algorithmes fondés sur le support et la confiance

Les algorithmes d'extractions des règles d'association fondées sur l'approche *support-confiance* procéder en deux phases [43] :

· ***Phase1*** : extraire les itemsets fréquents (dans un espace de recherche de taille exponentielle)
· ***Phase2*** : génération des règles d'association à partir des itemsets généré en phase1.

Vu que la phase 1 est la plus compliquée, ces algorithmes ont été classés en quatre classes, selon les stratégies de parcours de treillis des itemsets et la manière de calcul de support [47]. La table Tab. II.5 illustre cette classification.

Parcours en largeur		**Parcours en profondeur**	
Parcours de la base	Intersection des *tidlists*	Parcours de la base	Intersection des *tidlists*
AIS et Apriori	*Partition*	*FP-Growth*	*Eclat*

Table II.5 : Classification des algorithmes basés sur les mesures *Support-Confiance* [47]

A. L'algorithme AIS :

L'algorithme AIS tire son nom des initiales de ses concepteurs, Agrawal, Imielinski et Swami [43]. Il s'agit de l'un des premiers algorithmes dédiés à l'extraction de règles d'association. Son principe repose sur la génération de règles de la forme A → B, où A est un seul item, tandis que B est un ensemble d'items (*itemset*). Cet algorithme permet ainsi d'identifier des relations fréquentes entre les éléments d'une base de données transactionnelle, en mettant en évidence les associations significatives entre les items.

B. L'algorithme Apriori :

Paru en 1994 [54], Apriori est considéré comme l'algorithme de référence pour l'extraction des RA dans les bases de données transactionnelles. Il améliore significativement les performances des approches antérieures, notamment l'algorithme AIS, en optimisant le processus de génération des *itemsets* candidats. Son efficacité repose sur une stratégie d'élagage plus performante de l'espace de recherche, permettant de réduire considérablement le nombre de combinaisons à évaluer. Cette stratégie est basé sur le théorème de l'anti-monotonicité, selon lequel si un *itemset* est fréquent, alors tous ses sous-itemsets le sont également ; inversement, si un *itemset* est non fréquent, aucun de ses sur-itemsets ne peut être fréquent. Grâce à ce principe, Apriori élimine précocement de nombreux candidats non pertinents, ce qui améliore l'efficacité globale de l'algorithme et réduit le temps de calcul (voir la section II.7).

C. L'algorithme Partition :

Inspiré par [55] de l'algorithme *Apriori*, cet algorithme utilise une stratégie différente pour calculer les supports des itemsets candidats, en mémorisant avec chaque itemset fréquent A la liste de toutes les transactions qui le supportent, il stocke juste l'identifiant de la transaction qu'on a appellera *tidlist*. La génération des itemsets candidats est identique à celle d'*Apriori*, dont le support de chaque *Itemset* candidat est calculé via l'intersection des *tidlists* de tous ses items. Il faut noter que *Partition* divise la base de données en tranches adéquates à la taille de la mémoire physique.

D. L'algorithme Eclat :

Se base sur la notion de « classe d'équivalence » pour chercher les *Itemsets* fréquents en profondeur, et sur le format vertical de la base de données pour calculer les supports de ces *Itemsets* [56]. Ce calcul s'effectue grâce aux intersections des *tidlists*. Ceci réduit la taille de la base de données chargée dans la mémoire physique, car seules les transactions concernant un *Itemset* sont chargées et utilisées pour l'intersection.

E. L'algorithme FP-Growth (Frequent Pattern Growth) :

C'est un algorithme d'extraction de patterns fréquents utilisé pour générer des règles d'association [57]. En s'appuyant sur une structure d'arbre, *FP-Growth* parcours la base de données en deux étapes : la construction de l'arbre appelée FP-Tree (*Frequent-Pattern-Tree*) qui est une structure de donnée compacte de la base de données, ensuite la génération des patterns fréquents à partir de cette structure.

II.7 Algorithme Apriori pour l'Extraction des règles d'association

Dans cette section, on se focalise sur l'algorithme Apriori parce qu'il est la référence dans ce domaine. En plus, il est la base de beaucoup d'autres algorithmes d'extraction des règles d'association, aussi notre approche proposée est fondée sur cet algorithme (voir chapitre 4). L'algorithme *Apriori* et paru en 1994 [54]. *Apriori* est plus efficace que son algorithme prédécesseur AIS au niveau de la génération des *Itemsets* candidats, parce qu'il élague plus efficacement l'espace de recherche. Ceci résulte du théorème de l'anti-monotonie du support, qui est l'idée fondamentale de cet algorithme. En effet, *Agrawal* et *Srikant* [54] ont montré, que si un Itemset est fréquent alors tous les sous-ensembles de l'Itemset sont aussi fréquents, et inversement, tous les super-ensembles d'un itemset non fréquent sont obligatoirement non fréquents. La propriété d'anti-monotonie permet ainsi de minimiser le temps d'exécution et l'espace de recherche. Le nom de l'algorithme *Apriori* vient donc du fait qu'il tient en compte de la connaissance *antérieure* des itemsets fréquents.

L'algorithme Apriori fournit un ensemble de règles basées sur deux mesures statistiques, le support et la confiance de la règle. Il a apporté une solution élégante au problème du nombre important de règles extraites, basée sur le principe d'anti-monotonie, et en sélectionnant les itemsets et les règles qui ont un support et une confiance supérieurs respectivement aux deux seuils *Min_Sup* et *Min_Conf* définis par l'utilisateur, pour ignorer les itemsets non fréquents et les règles inutiles. La découverte de règles d'association par l'algorithme Apriori se fait en deux étapes principales : recherche des itemsets fréquents, génération des règles d'association utiles.

II.7.1 Recherche des itemsets fréquents

C'est la "partie principale" de l'algorithme Apriori. Initialement, il faut fixer un seuil *Min_Sup* pour que seuls les itemsets ayant un support ≥ *Min_Sup* soient générés. La phase de génération se déroule en deux étapes, dans la première étape (étape de *jointure*), les itemsets *candidats* de taille k sont générés en faisant une jointure des itemsets de taille k-1. Ensuite, dans la deuxième étape (étape d'*élagage*), tout itemset X généré par la phase de jointure est supprimé s'il existe un sous-ensemble de X qui est non fréquent (propriété d'anti-monotonicité). C'est un algorithme par niveau, il extrait à chaque niveau k les itemsets fréquents à partir combinaisons d'itemsets possibles.

Le pseudocode de l'algorithme Apriori est présenté dans l'algorithme 1 :

Algorithme 1 : Algorithme Apriori

Données : T : base de transactions, *Min_Sup* : support minimum
Sorties : LIF : Liste des Itemsets Frequents

```
1:  L1 = Liste des 1-itemsets frequents
2:  k = 2
3:  while Lk-1 ≠ ∅ do
4:      Ck = Genere_Candidats (Lk-1)          % génération des itemsets candidats
5:      for transaction t ∈ T do
6:          Ct = Subset(Ck, t)                % les candidats contenus dans Candidats[k]
7:              for candidat c dans Ct do
8:                  c.compteur++ ;
9:              end for
10:     end for
11:     Lk = {c ∈ Ck / c.compteur ≥ Min_Sup}
12:     k++
13: end while
14: LIF =Uk Lk                                % tous les itemsets fréquents trouvés
```

A chaque itemset, on associe un compteur noté c.compteur, pour stocker son support. Après, les 1-itemsets sont déterminés lors de la première passe sur la base de transaction T, en comptant les nombres d'apparitions de chaque item afin de déterminer ceux qui sont fréquents (L_1). Ensuite, la k^{ime} passe ($k \geq 2$) est composée de trois phases :

1. Les (k-1)-itemsets fréquents L_{k-1} trouvés lors de la k-$1^{iéme}$ passe sont utilisés pour générer les itemsets candidats de taille k, cette tâche est assurée par la fonction "Genere_Candidats (L_{k-1})"

2. A chaque fois, la table T est parcourue afin de calculer la fréquence des candidats, en utilisant la fonction "Subset" qui fournit les itemsets C_t contenus dans une transaction t.

3. On sélectionne les itemsets qui ont une fréquence supérieure ou égale au seuil *Min_Sup*.

L'algorithme Apriori s'arrête quand il n'y a aucun nouveau candidat de taille supérieure à générer. Le pseudocode de la procédure "Genere_Candidats (L_{k-1})" est le suivant :

Algorithme 2 : Algorithme de la procédure (Genere_Candidats (L_{k-1}))

```
1:  for itemset X ∈ L_{k-1} do
2:      for itemset Y ∈ L_{k-1} do
3:          if (X[1]=Y[1])^(X[2]=Y[2])^...^(X[k−2] =Y[k−2])^(X[k−1]<Y[k−1]) then
                                        % X<Y, X et Y partagent leur k-2 premiers items
4:              c = X[1], …, X[k-1], Y[k-1] % étape de jointure : générer des candidats
5:              Insérer c dans C_k
6:          end if
7:          for all (k-1)-sous-ensembles s de c do          % étape d'élagage
8:              if s ∉ L_{K-1} then
9:                  Supprimer c de C_k
10:             end if
11:         end for
12:     end for
13: end for
14: Retourner C_k
```

Comme résultats des algorithme 1 & 2, les itemsets fréquents ont les propriétés suivantes :

- ***Propriété 1 :*** Support pour les sous-itemsets :
  Pour les itemsets A et B, Si $A \subseteq B$ alors $supp(A) \geq supp(B)$,
  Car toutes les transactions qui supportent B supportent aussi nécessairement A.
- ***Propriété 2 :*** Les sous-itemsets d'itemsets fréquents sont fréquents.
- ***Propriété 3 :*** Les sous-itemsets d'itemsets non fréquents sont non fréquents

II.7.2 Génération des règles d'association

A partir de la liste des itemsets fréquents LIF trouvés dans l'étape précédente, une liste de règles d'association LR sera générée. Une règle est retenue si et seulement si sa confiance ≥ *Min_Conf* définie par l'utilisateur,

Agrawal et al. [54] ont proposé l'algorithme suivant (Algorithme 3) qui permet de générer ces règles, son pseudocode est le suivant :

Algorithme 3 : Génération des règles d'association

Données : LIF : Liste d'Itemsets Fréquents, *Min_Conf* : seuil minimal de confiance
Sorties : LR : liste des règles d'association générées

1: **for all** k-itemset frequent $L_k \in LIF$ tel que $k \geq 2$ **do**
2: Génèrer tous les sous-itemsets S_m de L_k
3: **for all** S_m de L_k **do**
4: $Conf(R) = Supp(L_k)/Supp(L_k - S_m)$
5: **if** $Conf(R) \geq Min\_Conf$ **then**
6: $LR \leftarrow LR \cup R \quad : \quad (L_k - S_m) \rightarrow S_m$
7: **end if**
8: **end for**
9: **end for**
10: Retourner LR

Pour chaque itemset fréquent L_k de taille > 1, on génère les S_m ($m < k$) qui sont des sous-ensembles de L_k. Pour chaque S_m on calcule la confiance *Conf* de la règle R : $(L_k - S_m) \rightarrow S_m$, si *Conf* ≥ *Min_Conf* (seuil minimal de confiance) on ajoute la règle R à LR (liste des RA générées). Comme résultats de l'algorithme 3, les règles extraites ont les propriétés suivantes :

- ***Propriété 4*** : Pas de composition des règles :
  Si $(X \rightarrow Z) \in LR$ et $(Y \rightarrow Z) \in LR$, $(X \cup Y \rightarrow Z)$ n'est pas nécessairement dans LR.
- ***Propriété* 5** : Décomposition des règles : si $X \cup Y \rightarrow Z$ est vrai, $X \rightarrow Z$ et $Y \rightarrow Z$ peut être faux
- ***Propriété* 6** : Pas de transitivité : si $X \rightarrow Y$ et $Y \rightarrow Z$, nous ne pouvons pas en déduire que $X \rightarrow Z$
- ***Propriété* 7** : Déduire si une règle convient :
  Si $A \rightarrow (L-A) \notin LR$ alors $B \rightarrow (L-B) \notin LR$ pour les itemsets L, A, B et $B \subseteq A$.

II.7.3 Exemple illustratif sur le fonctionnement de l'algorithme Apriori

Considérons les transactions T_i (les paniers d'achats des clients) dans d'une supérette qui vend 5 types de produits (5 items) : A, B, C, D, E. Chaque T_i représente les achats dans le jour i.

T_i	Items				
100	A	B	C	D	-
200	-	B	C	-	E
300	A	B	C	-	E
400	-	B	-	-	E

Table II.6 : Exemple d'un ensemble de transactions

Comme hyper-paramètres de cet algorithme, nous fixons les seuils minimums du support et de la confiance à : *Min_Sup* = 2 (donc 50% puisqu'on a 4 transactions) et *Min_Conf* = 70%. Le déroulement des étapes de l'algorithme est schématisé dans la figure Fig. II.8 suivante :

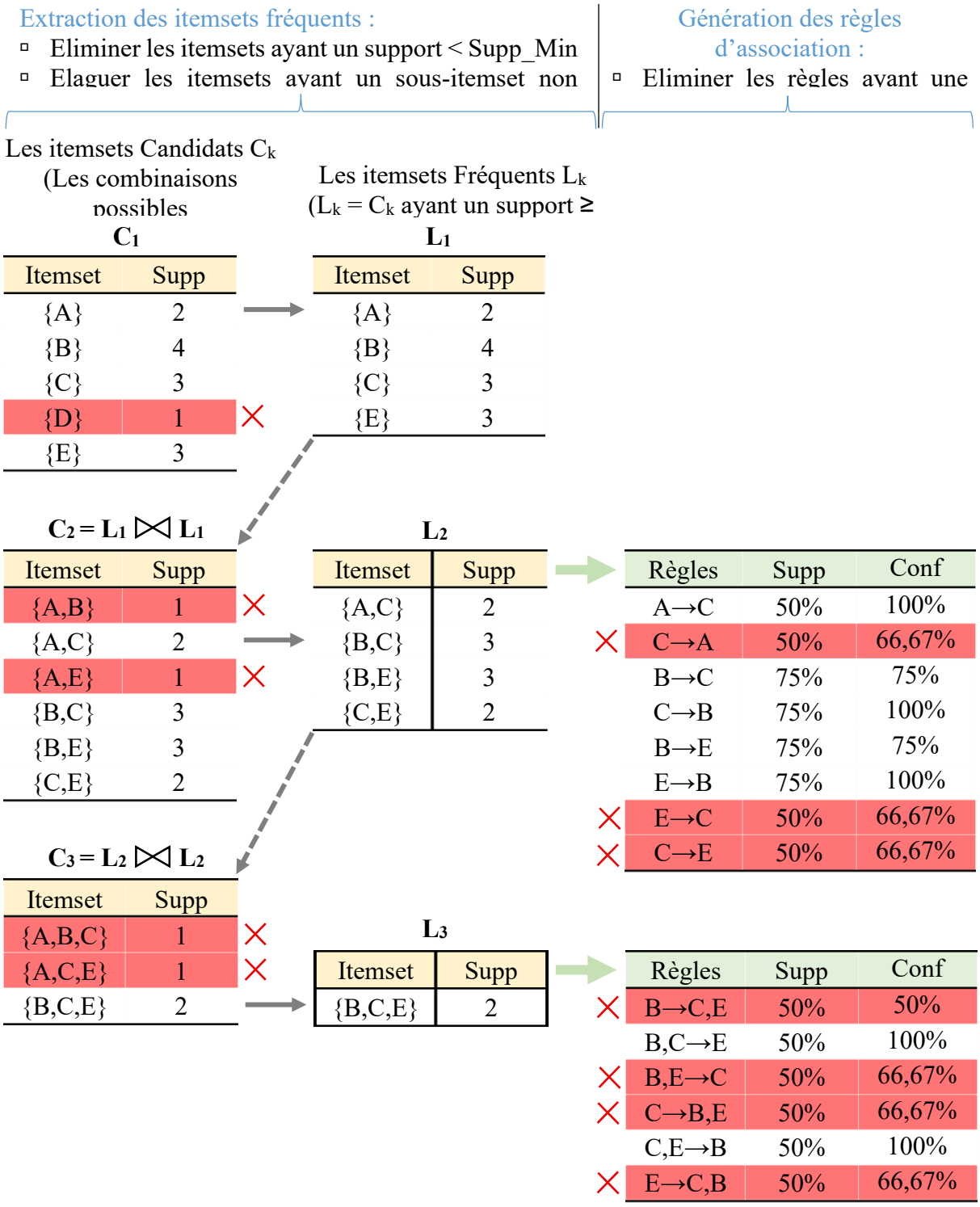


Figure II.8 : Exemple d'extraction des itemsets fréquents via Apriori et de génération des RA

Résultats : l'ensemble des règles d'association extraites à la fin de l'algorithme Apriori, présentées sous le format [Règle, Supp, Conf], sont comme suite :

{ [A→C, 50%, 100%], [B→C, 75%, 75%], [C→B, 75%, 100%], [B→E, 75%, 75%], [E→B, 75%, 100%], [B→CE, 50%, 100%], [CE→B, 50%, 100%] }

II.8 Les avantages et inconvénients de l'algorithme Apriori

Parmi les avantages des algorithmes d'extraction des règles d'association en général, et donc de l'Apriori, on peut citer :

- Les résultats des règles d'association sont faciles à interpréter et à comprendre, leur niveau explicatif ou sémantique est très élevé par rapport aux autres techniques comme les réseaux de neurones, par exemple, qui sont qualifiés comme boites noires (niveau explicatif faible).
- Cette méthode permette de découvrir des relations intéressantes entre les attributs dans des bases de données volumineuses.
- Aucune hypothèse préalable est exigée, cet algorithme est une forme d'apprentissage non supervisé, appliqué sur des données structurées sous forme de tables de transactions.
- Les règles d'association sont applicables fiablement dans divers domaines de recherche.
- Pour l'Apriori, le principe d'anti-monotonie du support aide à réduire la complexité de cet algorithme en réduisant l'espace de recherche des itemsets candidats.
- Cette méthode est facilement adaptable pour traiter d'autres types de données, comme :
  - Les séries temporelles : si un client achète un computer alors il achètera une imprimante dans les 3 mois suivant. Pour l'Apriori, une extension appelée AprioriAll était proposée pour ce type de règles.
  - Initialement, la quantité d'un produit acheté n'est pas prise en compte, la présence d'un produit dans une transaction exprime son achat, sans indiquer à quelle quantité il a été acheté. L'Apriori a été étendu pour traiter ce type de données en proposant les règles d'associations quantitatives.

Mais aussi, comme inconvénients de cet algorithme, on peut dire que :

- Son inconvénient majeur est sa *complexité algorithmique d'ordre exponentiel* :
  - L'algorithme nécessite un parcours répétitif de la base de données lors de chaque calcul du support.
  - A chaque itération on obtient un nombre énorme d'itemsets candidats. Pour un Itemset fréquent de taille n, l'algorithme génère au total $(2^n - 2)$ règles d'association possibles.
  - Aussi, on peut générer beaucoup de règles triviales (évidentes) ou inutiles (difficiles à interpréter).
  - Cet algorithme est donc très gourmand en temps de calcul (besoin d'un CPU puissant), et en espace mémoire (besoin d'une mémoire centrage vaste).
- Bien que, en Apriori par exemple, la contrainte du support minimum permet de diminuer le temps de calcul, cette contrainte peut aussi éliminer des règles rares mais très utiles.

II.9 Domaines d'application des règles d'association

Depuis la découverte de la méthode des règles d'association classiques (dites binaires) par Agrawal et al. [43], elle a montré son efficacité comme un outil d'aide à la décision à partir des bases de données. Elle a été largement utilisée dans nombreux domaines notamment :

- Le secteur commercial, selon [59] [60] cette technique permet de faire progresser les ventes en approchant l'emplacement des produits vendus fréquemment ensemble, mettre en promotion les produits non fréquents, identifier les préférences de différents groupes de clients et anticiper leurs besoins futurs, . . . etc.
- Dans le domaine de la santé, la découverte d'associations entre les symptômes des maladies peut aider les médecins au diagnostic de quelques maladies et prédire de traitement adéquat, et même de prévoir les symptômes futurs possibles du patient en fonction de ses antécédents, ou même à détecter les facteurs qui contribuent à la dégradation d'une maladie.
- En biologie, quelques phénomènes peuvent aussi se modéliser via une approche basée sur les règles d'association, comme la structure fonctionnelle des protéines qui est en forte association avec le séquencement des acides aminés qui composent cette protéine. Aussi, cet algorithme a été utilisé pour détecter des facteurs génétiques et environnementaux liés à une maladie.
- En WebMining, cette méthode est aussi appliquée pour : prédire le comportement successif des internautes qui consultent les pages d'un site web, étudier l'historique des ventes en ligne pour aider à améliorer la conception et l'organisation des sites d'e-commerce, déterminer les relations possibles entre les pages consultées afin de permettre aux internautes de trouver plus rapidement le contenu recherché sur un site web.
- En industrie, les règles d'association sont appliquées avec succès dans la résolution de plusieurs problèmes. En maintenance industrielle par exemple, cet algorithme est utilisé pour étudier la liaison entre les symptômes de pannes des machines de production, afin de prévoir les pannes avant qu'elles ne surviennent, et favoriser donc la maintenance préventive plus que corrective, ce qui assure la continuité des lignes de production.
- Aussi, les règles d'association sont appliquées pour prédire des phénomènes naturels et météorologiques, l'analyse de données de recensements, l'éducation et le E-learning…etc

II.10 Les axes d'extension des algorithmes d'extraction des RA

Les algorithmes présentés dans les sections précédentes visent la recherche des règles d'associations dites "binaires". Plusieurs extensions ont été proposées, qui sont des améliorations et adaptations des algorithmes classiques de recherche d'associations, pour étendre l'analyse à d'autres types de données et à d'autres types de logiques de modélisation. Les variantes déjà vues dans la section II.6 (C5.0, GRI, AIS, Apriori, Partition, Eclat, FP-growth, …) sont basées notamment sur l'amélioration des performances fonctionnelles de l'algorithme d'extraction des règles d'association, comme : la réduction de l'espace de recherche pour éviter l'explosion combinatoire des itemsets possibles, ou bien la proposition de métriques adéquates pour filtrer les règles les plus pertinentes. Mais ces améliorations ont toujours considéré les données analysées comme certaines et en format binaire. En ce qui suit, nous allons présenter d'autres extensions de cet algorithme qui ont été proposées pour élargir le spectre du type de connaissances qu'on peut extraire à partir des datasets analysés.

II.10.1 Extensions basées sur la nature des règles extraites

Suivant le type de données à analyser et afin d'extraire d'autres types de règles d'association, plusieurs extensions de cet algorithme on été developpées, telles que : les règles d'association quantitatives et généralisées, ainsi que les algorithmes d'extraction des motifs séquentiels.

II.10.1.1 Les règles d'association quantitatives

Parmi les points faibles fondamentaux des algorithmes basiques on trouve le format binaire, utilisé dans la représentation des transactions. Par exemple, si dans les transactions on signale l'achat d'un produit, on ne fait aucune différence entre les quantités achetées de ce produit dans chaque transaction. Par exemple, si un produit (item) B est acheté à une quantité faible (par exemple 1 ou 2) en achetant le produit A, cela sera la même chose si un autre produit C est acheté à une quantité élevée (exemple 50 ou 60) si le produit A est acheté. Les relations A→B et A→C peuvent avoir la même confiance et seront donc considérées comme équivalentes, mais la relation d'achat entre A et C est bien sûr la plus intéressante.

Cette connaissance, très utile pour augmenter les achats, sera masquée (et donc ignorée) dans les algorithmes d'extraction des règles d'association binaires. Les règles d'association quantitatives [61] sont une approche basée sur les règles d'association binaires, aidant à extraire des relations de la forme suivante : « Si (Age est [18 ... 35]) alors (salaire est 18000 DA) ». Cette variante peut analyser des données « qualitatives » (exemple : couleurs (bleu, rouge…)), et des données « quantitatives » (comme les mesures M, Kg, etc.). L'idée est que les attributs quantitatifs peuvent être subdivisés en ensembles d'intervalles, de même que toutes les valeurs d'un attribut qualitatif peuvent être remplacées par des valeurs entières.

II.10.1.2 Les règles d'association généralisées

Introduites par [62], les règles extraites via cet algorithme sont similaires aux règles d'associations binaires. La différence principale consiste à prendre en compte une éventuelle structure hiérarchique appelée « taxonomies » sur l'ensemble des Items. L'idée de base des règles d'association généralisées est que les articles dans les grands magasins sont décrits par un code détaillé. Par exemple, des boots de taille 29 à 35 ont un code, et des boots de taille 36 à 39 ont un code différent. Si on cherche des règles d'association ayant un support suffisant, ces codes vont diminuer les chances de les trouver. Mais, si on arrive à indiquer à l'algorithme que les boots de tailles différentes sont tous des boots, alors le support de l'article *boots* sera la somme des supports des boots de différentes tailles. Ainsi, on augmente les chances de produire des règles d'association sur les *boots* avec un support suffisant.

II.10.1.3 Les motifs séquentiels

Les motifs séquentiels sont introduits dans [63] comme une extension de l'algorithme d'extraction des règles d'association. Les auteurs ont proposé l'algorithme AprioriAll, qui est une adaptation de l'algorithme Apriori, adéquate à l'extraction des relations sous forme de séquences temporelles dans les bases de transactions. En effet la prise en compte de la temporalité (l'attribut *date)* dans les enregistrements à étudier permet une plus grande précision dans les résultats, mais implique aussi un plus grand nombre de calculs et de contraintes. Le problème d'extraction des motifs séquentiels est de trouver les séquences maximales parmi toutes les séquences qui ont un support supérieur à un *MinSup* spécifié par l'utilisateur. Dans cet algorithme, une transaction constitue, pour un client C, l'ensemble des items achetés par C à une même date, une séquence est donc une suite de transactions qui apporte une relation d'ordre (enchaînement temporel) assez fréquente entre les transactions.

II.10.2 Extensions basées sur la logique de modelisation

Comme nous l'avons cité auparavant, les règles d'associations ont été conçues initialement pour traiter les données au format binaire, et donc en logique qui modélise seulement les relations strictes entre les attributs d'une base de données, mais le monde est loin d'être en logique binaire.

Puisque dans la réalité, on peut traiter des données incertaines ou imprécises, et dans le but d'extraire des règles d'association en adéquation avec la réalité des données étudiées, qui peuvent ne pas être binaires, des adaptations de cet algorithme ont été proposées pour extraire des règles d'association en adéquation avec des logiques de modélisation plus avancées que la logique : floue, possibiliste, évidentielle.

II.10.2.1 Les règles d'association floues

Issues de l'introduction de la théorie des sous-ensembles flous [64] dans la recherche des relations entre les items [67], cette hybridation a permis de modéliser l'incertitude et l'imprécision des données étudiées. Contrairement aux règles d'association classiques, qui reposent sur des relations strictes entre les attributs, les règles d'associations floues prennent en compte le degré d'appartenance d'un élément (item) à un ensemble (sous-ensemble flou) via une fonction d'appartenance.

L'un des principaux avantages des algorithmes de Machine Learning modélisés en logique floue est leur capacité à traiter des informations incertaines et à fournir des connaissances plus nuancées par rapport aux méthodes classiques. Ainsi, les règles d'associations floues permettent une analyse plus flexible et réaliste des données étudiées, en prenant en compte les incertitudes et les nuances qui peuvent exister dans les informations disponibles. On trouve dans la littérature plusieurs travaux qui ont proposé des règles d'associations floues pour l'extraction de connaissances dans plusieurs domaines [68-70]. Plus de détails sur ces règles seront fournis dans le chapitre suivant.

II.10.2.2 Les règles d'association possibilistes

Comme extension de la logique floue, la logique possibiliste [71] est également utilisée pour extraire des règles d'associations en adéquation avec la réalité nuancée des données étudiées. Des travaux de recherche comme [72] ont prouvé la pertinence de leurs résultats. Ainsi, les règles d'association floues et possibilistes traitent toutes deux de l'incertitude, mais les règles floues se concentrent sur le degré d'appartenance des items aux sous-ensembles flous, tandis que les règles possibilistes évaluent la plausibilité des associations entre les items.

Les règles d'associations possibilistes offrent un cadre puissant pour analyser des données incertaines et fournir des connaissances pertinentes dans des situations parfois plus complexes que celles modélisées via la logique floue, tout en tenant compte des nuances et des incertitudes inhérentes aux informations disponibles. Plus de détails sur ces règles seront également fournis dans le chapitre suivant.

II.10.2.3 Les règles d'association évidentielles

Les règles d'associations évidentielles sont basées sur la théorie des croyances (ou théorie de Dempster-Shafer) [73][74]. Cette approche permet de modéliser l'incertitude en se basant non seulement sur des données existantes, mais aussi sur des notions de la masse de croyance et des informations incomplètes et/ou contradictoires [75]. La Théorie Dempster-Shafer permet de représenter l'incertitude en utilisant des plages de croyance plutôt que des probabilités précises. Elle repose sur la notion de masse de croyance, qui quantifie le degré de soutien d'une hypothèse donnée.

Les règles d'associations évidentielles offrent un cadre robuste pour analyser des données incertaines et fusionner des informations provenant de différentes sources, en tenant compte

des croyances et des incertitudes inhérentes. Cette approche est particulièrement adaptée aux situations complexes où les données sont conflictuelles ou ambiguës. Cette logique de modélisation sera aussi détaillée dans le chapitre prochain.

II.11 Conclusion

Les règles d'association sont l'une des techniques les plus utilisées pour extraire les connaissances cachées dans les bases de données transactionnelles, vu la qualité des connaissances extraites et la facilité de leur interprétation. Cet algorithme a montré son utilité dans plusieurs domaines d'applications pour découvrir des relations cachées et des associations inattendues entre les attributs d'une base de données. L'algorithme Apriori est la référence dans cette technique, il est la base de plusieurs algorithmes développés.

Cependant, les algorithmes classiques d'extraction des RA, y compris l'Apriori, analysent des données supposées exactes et formatées en logique binaire. Mais, dans le monde réel, on sera souvent face à des données de nature ambiguës (incertaines et/ou imprécises), et donc l'algorithme Apriori doit être adapté à des logiques de modélisation plus adéquates à la réalité des données étudiées (qui sont généralement loin d'être binaires). Pour cela, dans le chapitre suivant, nous présentons les concepts de base des logiques évoluées de modélisation (logiques : floue, possibiliste, évidentielle), et les méthodes d'adaptation des algorithmes d'extraction des règles d'association à ces logiques.

Chapitre 3

Chapitre III : Adaptation des règles d'association aux théories de l'incertain

III.1 Introduction

Beaucoup de connaissances, généralement inexploitées par les biologistes, sur le développement de l'embryon sont cachées dans des « séquences d'images » représentant « les zones d'expressions génétiques » de chaque gène dans les « phases de croissance » de cet embryon. On parle ainsi de bases de données spatiotemporelles, contenant un grand volume de données qui cachent beaucoup de connaissances cruciales et inconnues. Ces connaissances aident les biologistes à suivre et mieux comprendre les interactions génétiques lors des phases de développement des embryons de l'espèce modèle « Edinburgh Mouse », et donc des espèces vivantes, et à étudier l'impact des expressions génétiques dans chaque phase.

Ainsi, introduire les théories de modélisation sous l'incertain (comme la logique des sous-ensembles flous, possibiliste, évidentielle) pour proposer une représentation formelle des données biologiques qui ont parfois un aspect imprécis et donc une signification incertaine, sans avoir l'obligation qu'une vérité biologique doit être « totalement » ou « nullement » absolue, est plus fidèle à la complexité profonde du génome des vivants qui cache toujours des secrets et révèle à chaque fois les miracles du *grand créateur*. Les théories de l'incertain, qu'on va détailler dans ce chapitre, ont montré de très bons résultats en grande adéquation avec la réalité des données étudiées dans les différents domaines d'application, notamment celui de l'approche proposée dans cette thèse.

Partie 1 : Théories de l'incertain

III.2 De la logique binaire aux logiques de l'incertain

III.2.1 Le développement de la logique humaine

Dès l'aube de la philosophie, la connaissance était un sujet d'études privilégié. Les débuts du « raisonnement scientifique » étaient dans la « Grèce antique » et remontent à l'ère de Socrate, Platon et Aristote [76] :

- ***Socrate (l'éducateur) (470-399 Avant Jésus-Christ)*** *:* il enseignait dans les rues d'Athènes pour rendre les gens plus sages par la connaissance de leur ignorance. Il disait : *« Je sais que je ne sais rien ».*
- ***Platon (le philosophe) (428-348 av. J-C) :*** le père de la philosophie, Platon inspirait son raisonnement des autres sciences comme les théories mathématiques développées à l'époque par Thalès (625-547 av.J-C) et par Pythagore (580-495 av.J-C). Il a créé la première académie (vers 385 av.J-C) où une inscription au-dessus de son portail stipulait que « des connaissances en géométrie étaient une condition pour étudier la philosophie ».
- ***Aristote (le logicien) (384-322 av. J-C) :*** élève de Platon et enseignant d'Alexandre le grand, Aristote a créé le premier lycée. Il est considéré comme le « fondateur de la logique formelle ». Ensuite, d'une civilisation à une autre (babélienne, romaine, perse, chinoise, islamique, occidentale, ...) la logique humaine n'a cessé de se développer pour acquérir la connaissance.

III.2.2 La logique binaire

III.2.2.1 Algèbre de Boole

L'algèbre de Boole est une branche des mathématiques qui étudie les opérations appliquées à des variables logiques ne pouvant prendre que deux valeurs possibles : 0 (faux) et 1 (vrai). Elle tire son nom du mathématicien George Boole (1815-1864) [77], considéré comme l'un des fondateurs de la logique moderne, pour avoir formalisé le raisonnement logique sous la forme d'un système mathématique rigoureux. De nos jours, l'algèbre de Boole est largement utilisée en informatique, notamment dans la conception des circuits numériques et le traitement de l'information. Pour deux variables logiques A et B, les opérations fondamentales sont la conjonction ($A \cdot B$), la disjonction ($A + B$) et la négation ($\neg A$). Une valeur logique est ainsi représentée par un booléen. Il convient également de distinguer plusieurs niveaux de calcul logique, tels que le calcul booléen, le calcul des équations booléennes correspondant au calcul propositionnel, le calcul propositionnel d'ordre supérieur et le calcul des prédicats. Les logiques binaires se déclinent en différentes catégories, comme les logiques d'ordre 0, 1 et 2, ainsi que les logiques d'ordre supérieur, utilisées dans les langages de programmation fonctionnels [78].

III.2.2.2 Pourquoi le monde réel n'est pas en logique binaire

La logique classique utilise un système de représentation binaire, ce qui signifie qu'une proposition peut être soit vraie, soit fausse (0 ou 1). Cependant, ce système binaire ne convient pas à tous les problèmes rencontrés. Dans la vie quotidienne, l'individu observe, conçoit, raisonne et fait des choix basés sur des modèles ou des représentations. Néanmoins, il est évident que la pensée humaine ne se limite pas à une approche binaire ; elle ne s'apparente pas à une science exacte, et les décisions concernant des situations courantes ne se fondent pas uniquement sur des valeurs de vérité ou de fausse [79]. Quand on prend un exemple du monde réel, on trouve que la logique binaire partitionne la température pour un certain jour en deux états (chaud ou bien froid), mais dans la réalité il y a une phase entre les deux ; c'est la température modérée. En effet, on trouve le même raisonnement pour plusieurs phénomènes dans le monde réel, qui est plein de données imparfaites, et donc loin d'être en logique binaire.

III.2.3 Les types d'imperfection dans les données réelles

L'imperfection dans les données peut être classifiée en trois catégories (non exclusives) : l'*incertitude*, l'*inconsistance*, l'*incomplétude* et l'*imprécision* ; chacune pouvant se décliner en plusieurs sous catégories. La figure suivante illustre ces différents types d'imperfection :

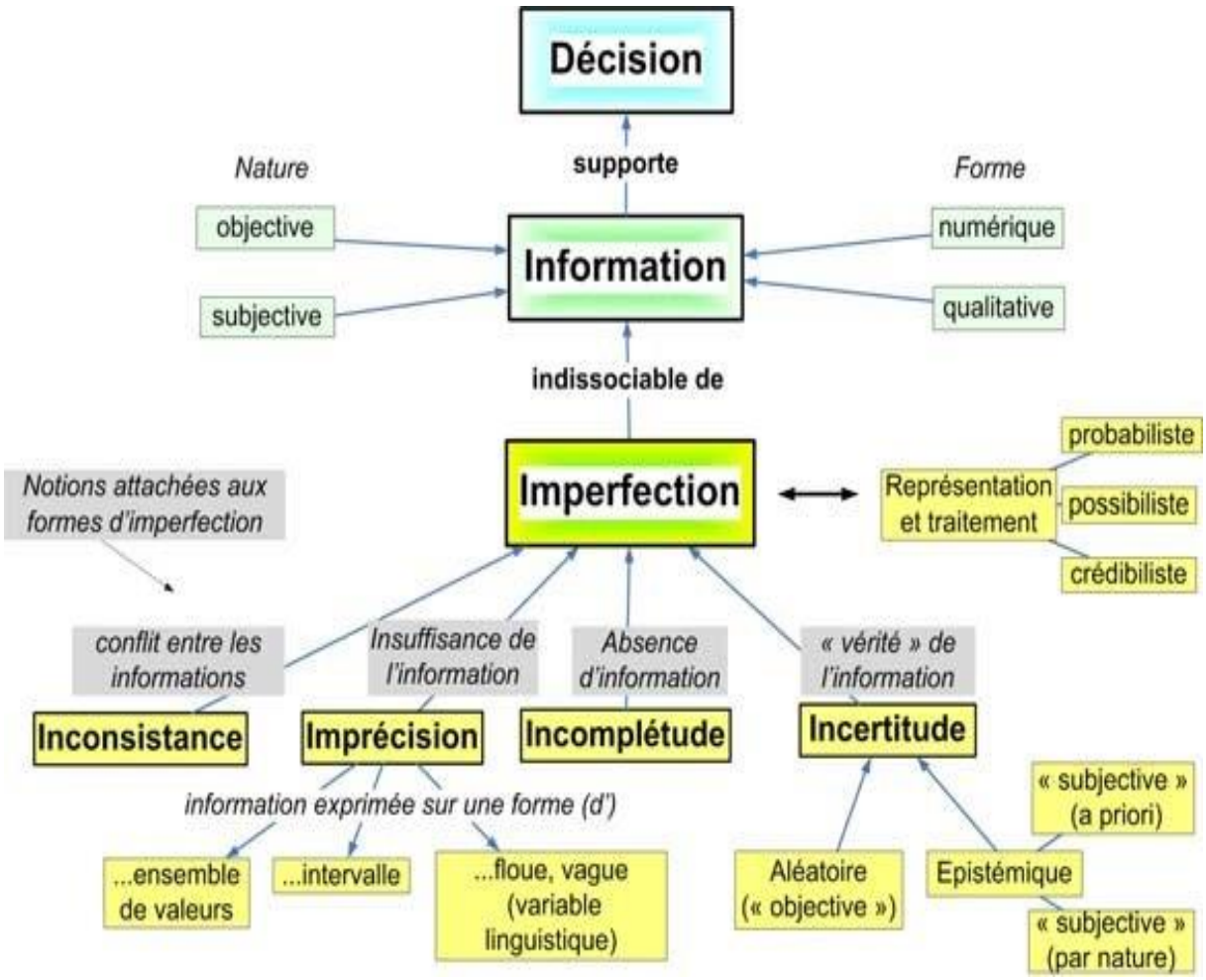


Figure III.1 : Les différentes formes d'imperfection des données [80]

III.2.3.1 Les données incertaines

L'incertitude se rapporte à la fiabilité des informations. En d'autres termes, elle implique un questionnement sur la validité d'une donnée. Par exemple, si un inconnu nous communique le résultat d'un match, il est possible que l'information soit complète et précise, mais néanmoins incorrecte. Certains auteurs identifient deux catégories [79] :

L'incertitude objective *:* on peut l'assimiler aux événements imprévisibles, les aléas de la vie !

L'incertitude subjective : est associée à la confiance que l'on accorde à la source d'information (par exemple, le supporter est conscient que la personne inconnue qui lui a donné le score n'est pas digne de confiance).

À ces deux concepts, on peut relier les idées d'incertitude :

Probabiliste : quelle est la probabilité de victoire de l'Algérie lors de son prochain match ?

Possibiliste : la probabilité que la différence de buts dépasse 5.

Crédibiliste : ma conviction personnelle que l'Algérie remportera le match.

III.2.3.2 Les données imprécises

Les connaissances numériques sont souvent mal comprises, car des termes du langage courant sont employés pour décrire une caractéristique du système de manière vague et imprécise. Typiquement la modélisation de l'imprécision contient des concepts vagues comme environ, proche, grand, … etc. En d'autres termes l'imprécision c'est une généralisation de l'incomplétude. Voici quelques exemples de l'imprécision [79] :

- L'équipe nationale va rencontrer au tour prochain soit la Tunisie ou le Sénégal (variable qualitative avec 2 modalités possibles).
- La différence entre le nombre de buts était inférieure à 3 (variable quantitative ayant une valeur appartenant à un intervalle).
- L'imprécision causée par une ambigüité linguistique (expression vague ou flou). Par exemple, l'Algérie a battu l'Égypte sur une différence de but "faible" (expression vague).

III.2.3.3 Les données incomplètes

Les incomplétudes se réfèrent à des manques de connaissances ou à des informations incomplètes concernant le système [81] :

Absence : l'absence se produit lorsqu'il y a des valeurs manquantes dans une base de données pour décrire certains objets. On évoque l'absence quand une information essentielle pour la définition complète d'un objet est manquante.

Lacune : la lacune désigne le cas où un ou plusieurs objets de la base de données fournissent uniquement une description partielle d'une structure qui les englobe.

Donc, l'incomplétude est l'absence d'information, exemple : résultat du match est inconnu.

III.2.3.4 Les données inconsistantes

L'inconsistance (incohérence) survient en présence de redondance et lorsque plusieurs informations sont en confit. On dit qu'il y a confit si au minimum deux classifications contradictoires pour le même objet sont possibles [79].

Exemple : Le match de l'équipe nationale a commencé il y a 5 min et le score est de (5 à 0)

III.3 Les logiques de modélisation dans l'incertain

III.3.1 La logique probabiliste et le théorème de Bayes

C'est la théorie la plus ancienne et la plus largement étudiée concernant l'imparfait. Elle trouve ses racines dans l'analyse des jeux de hasard. La probabilité d'un événement décrit la possibilité (le pourcentage de chances) que cet événement se produise. Considérant un univers **Ω**, qu'il soit discret ou continu, la probabilité P(A) d'un événement A ⊆ W se situe dans l'intervalle [0,1]. De plus, en prenant 2^{Ω} comme l'ensemble des parties de **Ω**, la probabilité est ainsi définie : la probabilité de l'événement **Ω** est égale à 1, tandis que celle de l'événement vide est de 0. En résumé, une probabilité est une fonction, notée P, qui associe à chaque événement A une valeur P(A) représentant la probabilité que A se réalise. Une probabilité présente les propriétés suivantes [81] :

- $0 \leq P(A) \leq 1$ pour tout évènement A.
- $P(\Omega) = 1$.
- $P(\emptyset) = 0$
- $P(\bar{A}) = 1 - P(A)$
- $A \subseteq B \Rightarrow P(A) \leq P(B)$

La probabilité et le raisonnement bayésien sont intimement liés, le raisonnement bayésien est une façon particulière d'utiliser les probabilités pour raisonner en présence d'incertitude. La probabilité sert à quantifier l'incertitude, tandis que le raisonnement bayésien utilise les probabilités pour mettre à jour une croyance quand on obtient une nouvelle information.

Notre cerveau est confronté au défi de l'incertitude, que ce soit lors de la perception de notre environnement physique, lors de la prise de décision ou au moment d'agir. Cette incertitude se manifeste à divers niveaux : au niveau du système sensoriel, de l'appareil moteur, de nos connaissances et du monde lui-même. La théorie probabiliste, et en particulier la théorie bayésienne, offre des outils pour gérer ces incertitudes. Le théorème de Bayes, qui constitue le fondement de cette théorie, indique, dans sa forme la plus simple, que notre perception du monde doit être ajustée en fonction du produit de nos croyances a priori et de nos observations. La formule de base du théorème de Bayes est [82] :

$$\mathbf{P(A / B) = P(B /A)\, P(A) / P(B)} \qquad \text{(III.1)}$$

P(A) = la "probabilité a priori"
(B/A) = la "vraisemblance" (ou probabilité conditionnelle) que l'événement B se produise étant donné que A est vrai ;
P(A/B) = la "probabilité a posteriori" ou encore la probabilité conditionnelle que l'événement A se produise étant donné que B est vrai ;
P(B) = la probabilité d'observer B.

Ainsi, l'inférence bayésienne, en se basant sur le théorème de Bayes, permet de déduire la probabilité d'un événement à partir de celles d'autres événements déjà évalués, autrement dit en fonction de l'historique des événements passés.

III.3.2 La logique floue

III.3.2.1 Généralités

La notion d'ensembles flous a été introduite par Lofti A. Zadeh en 1965 [64], basée sur l'idée d'appartenance partielle à une classe, généralisant ainsi la théorie classique des ensembles. Cela permet d'éviter les transitions abruptes d'une classe à une autre, comme celle du blanc au noir, et d'offrir une adhésion graduelle des éléments. Par exemple, un gris clair peut avoir un fort degré d'appartenance à la classe blanche et un faible degré à la classe noire. Ce degré d'appartenance détermine la position de l'élément au sein de l'ensemble flou. Par conséquent, un élément peut appartenir à plusieurs ensembles, ce qui est réalisé grâce à ces degrés d'appartenance. Ces derniers sont des nombres réels compris entre 0 et 1. Plus le degré d'appartenance est élevé, plus il se rapproche de 1, et inversement.

Dans la logique classique, l'appartenance d'un élément à un sous-ensemble donné est représentée par une valeur binaire (0 ou 1). La valeur 1 signifie que l'élément appartient totalement au sous-ensemble considéré, tandis qu'une valeur nulle indique que l'élément n'appartient pas à ce sous-ensemble. Le concept d'ensemble flou permet d'aborder [83] :

- Des classes aux frontières mal définies (catégories d'appréciation perçues par un observateur)
- Des classes intermédiaires entre le tout et le rien (par exemple, "presque certain")
- Le passage progressif d'une classe à une autre (comme du "petit" au "grand", ou du "faible" au "fort")
- Des valeurs approximatives (telles que "autour de 13 de moyenne" ou "environ 5m de distance").

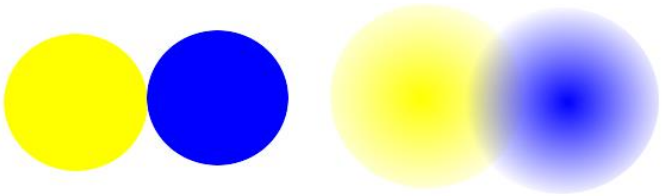


Figure III.2 : Logique classique (stricte) et logique floue (nuancée)

Exemple : " Il fait chaud aujourd'hui ". Chaud ?
Il existe toute une gamme de températures qui rend cette affirmation vraie :

- T=40°C : Certainement !
- T=30°C : Oui

- T=20°C : Un peu
- T=10°C : Pas tant que ça !
- T= 0°C : Pour un ours polaire, peut-être !

III.3.2.2 Caractérisation des Sous-Ensembles Flous

Voici les caractéristiques de base d'un Sous-Ensemble Flou (SEF) [83] :

- ***Le noyau*** d'un SEF A de X, noté $Noy(A)$, est constitué de tous les éléments qui y appartiennent entièrement (avec un degré d'appartenance = 1) : $\boldsymbol{Noy(A) = \{x \in X, \mu A\ (x)=1\}}$
- ***Le support*** d'un SEF A de X, noté supp(A), correspond à l'ensemble de tous les éléments ayant un degré d'appartenance supérieur à zéro : $\boldsymbol{Supp\ (A) = \{x \in X, \mu A\ (x) \neq 0\}}$
- ***La hauteur*** d'un SEF A de X, notée h(A), correspond à la valeur maximale atteinte par sa fonction d'appartenance : $\boldsymbol{t : h\ (A) = x \in X, \mu A\ (x)}$
- ***La cardinalité*** d'un SEF A de X, notée |A|, est la somme des degrés d'appartenance des éléments de X qui sont inclus dans A. Elle est définie par : $\boldsymbol{|A| = \sum \mu A\ (x)}$
- ***Une α-coupe*** d'un SEF A de X est un sous-ensemble constitué des éléments ayant un degré d'appartenance ≥ seuil d'appartenance α : $\boldsymbol{\alpha\text{-}coupe(A) = \{x \in X, \mu A\ (x) \geq \alpha\}}$

III.3.2.3 Fonctions d'appartenance

Chaque SEF peut être décrit par sa fonction d'appartenance. En règle générale, la forme des fonctions d'appartenance varie en fonction de l'application. En considérant A comme un sous-ensemble classique et X comme un ensemble d'éléments, XA représente la fonction caractéristique. Cette fonction prend la valeur 0 pour les éléments X qui n'appartiennent pas à A et la valeur 1 pour ceux qui y appartiennent. Elle peut être exprimée par : XA → {0, 1}.

$$x_A(x) = \begin{cases} 1\ si\ x \in A \\ 0\ sinon \end{cases}$$

Un SEF A est caractérisé par une fonction d'appartenance qui attribue à chaque élément x de X un degré d'appartenance μα(x) variant entre 0 et 1 : μα(x) : X → [0, 1] :

$$x_A(x) = \begin{cases} \mu_\alpha(x)\ si\ x \in A \\ 0\ sinon \end{cases}$$

Elle peut être de forme triangulaire, trapézoïdale ou gaussienne [83] :

- Fonction d'appartenance triangulaire (générale et symétrique) ;
- Fonction d'appartenance trapézoïdale (à gauche et à droite) ;
- Fonction d'appartenance gaussienne (à gauche et à droite ou pseudo-exponentielle).

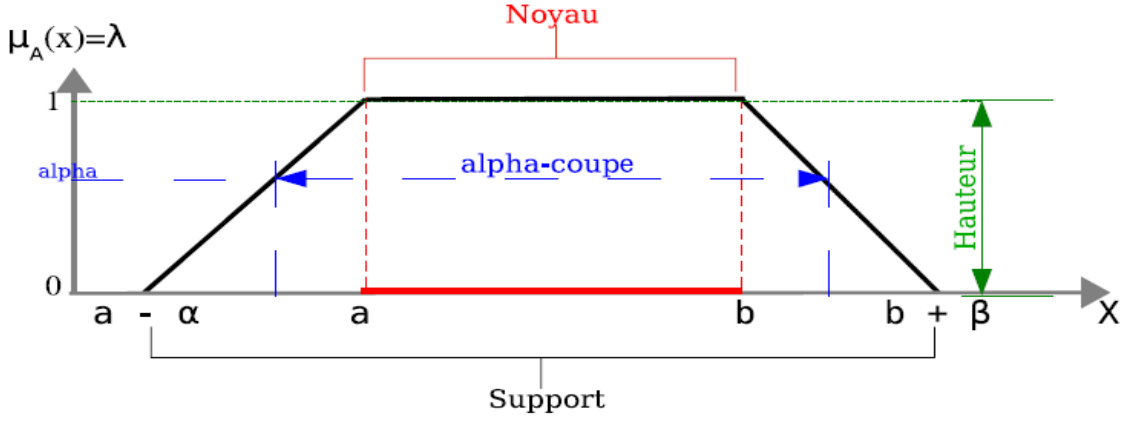
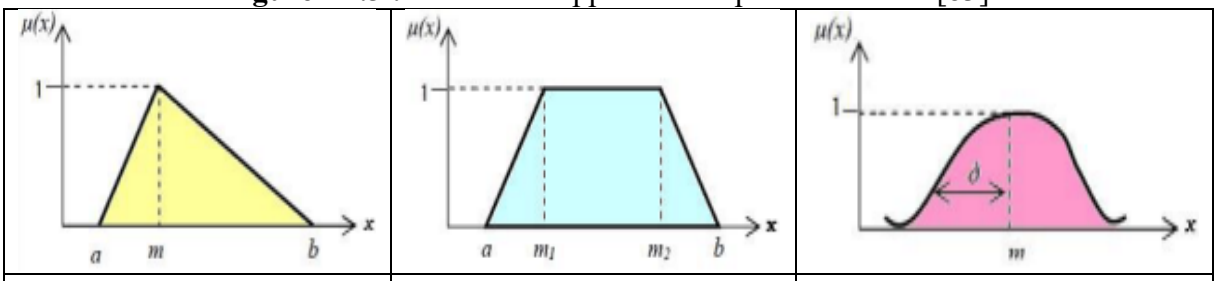


Figure III.3 : Fonction d'appartenance pour un SEF A [83]



Triangulaire	Trapézoïdale	Gaussienne

Figure III.4 : Fonctions d'appartenances (triangulaire, trapézoïdale, gaussienne)

III.3.3 La logique possibiliste

III.3.3.1 Les distributions de possibilité

La théorie des possibilités repose sur les distributions de possibilité [71] [84]. Considérant un univers de discours $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$, un concept clé noté π correspond à une fonction qui attribue à chaque élément ω_i une valeur dans un ensemble limité et ordonné de manière linéaire (L, <). Cette valeur, appelée degré de possibilité, représente les connaissances sur le monde réel. Par convention, $\pi(\omega_i) = 1$ indique qu'il est totalement possible que ω_i soit réel, tandis que $\pi(\omega_i) = 0$ indique que ω_i est impossible. La flexibilité est représentée en permettant l'attribution de valeurs dans l'intervalle]0, 1[. Dans la théorie des possibilités, les cas extrêmes sont modélisés par [85] :

- Connaissance complète : $\exists\omega_i \in \Omega$, $\pi(\omega_i) = 1$ et $\forall\omega_j \neq \omega_i\ \pi(\omega_j) = 0$;
- Ignorance partielle : $\forall\omega_i \in A \subseteq \Omega$, $\pi(\omega_i) = 1$ $\forall\omega_i \notin A$, $\pi(\omega_i) = 0$;
- Ignorance totale : $\forall\omega_i \in \Omega$, $\pi(\omega_i) = 1$.

III.3.3.2 Les mesures de possibilité et de nécessité

Une distribution de possibilité π sur Ω permet d'analyser les événements en fonction de leur plausibilité et de leur certitude, en utilisant deux mesures duales nommées Possibilité et Nécessité. Pour une distribution de possibilité π sur un univers de discours Ω, les valeurs de possibilité et de nécessité évaluent chaque événement $A \subseteq 2^{\Omega}$ de la manière suivante [85] :

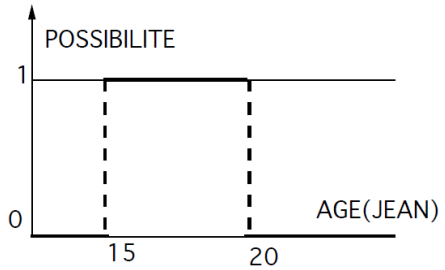
- La possibilité : $\Pi(A) = max_{\omega \in A}\pi(\omega)$
- La nécessité : $N(A) = min_{\omega \notin A}(1 - \pi(\omega)) = 1 - \Pi(\bar{A})$

$\Pi(A)$ mesure dans quelle mesure A est cohérent avec nos connaissances exprimées par π, alors que $N(A)$ évalue le degré de certitude de A selon nos connaissances. La taille du gap entre $N(A)$ et $\Pi(A)$ indique le taux d'ignorance sur A.

On peut distinguer deux cadres principaux pour représenter l'information imprécise. Le premier utilise des *intervalles*, qui permettent d'exprimer une valeur comme appartenant à un

ensemble de valeurs possibles, utilisé lorsque l'imprécision provient d'erreurs de mesure ou d'une connaissance partielle des données. Le second cadre est celui de la *logique symbolique*, qui modélise l'imprécision à l'aide de règles et de degrés de vérité, comme dans la logique floue ou possibiliste. Cette représentation permet de raisonner sur des notions vagues ou subjectives, telles que « grand », « faible » ou « proche », en se rapprochant du raisonnement humain [86].

A- **Les intervalles :** c'est la représentation de l'information imprécise par des ensembles, ici il ne faut pas confondre l'imprécision et les variables multivaluées. Ces ensembles sont formés d'éléments disjonctifs mutuellement exclusifs dont l'un d'eux est la vraie valeur d'une grandeur. Exemple : Age(JEAN) ∈ [20, 25] = 20 ou 21 ou 22 ou 23 ou 24 ou 25



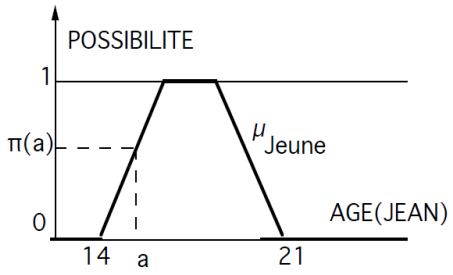
Exemple d'information imprécise :
JEAN a entre 15 et 20 ans

C'est une distribution de possibilité à valeurs dans {0, 1} : π(x) = 1 ssi x ∈ A et 0 sinon.

Fig. III.5 (A) : Représentation de l´imprécision via des *intervalles d'éléments disjonctifs*

B- **La logique symbolique :** comme la logique possibiliste, Une distribution de possibilité est une représentation d'un état de connaissance d'un agent relatif à l'état du monde.
U : référentiel (ensemble d'états du monde)
x : variable mal connue à valeur sur U
L : Echelle de plausibilité : ensemble totalement ordonné ([0,1], fini, ...)

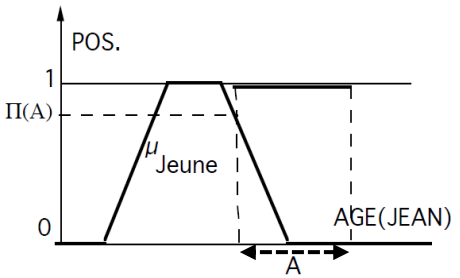
Une distribution de possibilité π_x attachée à x est une fonction de U dans L telle que ∀u, $\pi_x(u)$ ∈ L, et ∃u, $\pi_x(u) = 1$ (normalisation)



Exemple : « JEAN est Jeune »
$\mu_{Jeune}(x)$ = possibilité (AGE(JEAN) = x)

Jeune = sous-ensemble flou des valeurs possibles de l´AGE de JEAN
Cas particulier : JEAN a entre 15 et 20 ans :
possibilité(x) = 1 si 15 ≤ x≤ 20, et 0 sinon

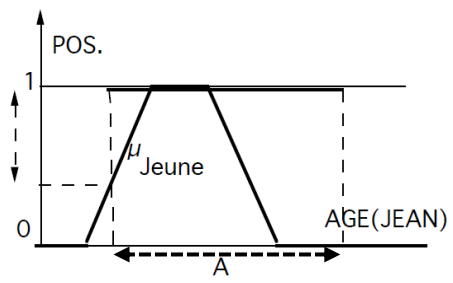
Fig. III.5 (B) : Représentation de l´imprécision en termes de *possibilité nuancée*



Π(A) = max(π(u) | u ∈ A) = Hauteur (A ∩ 'JEUNE')

Degré de cohérence entre une proposition A et l'information disponible = à quel point A n'est pas incompatible avec l'information "JEUNE"

Fig. III.5 (C) : Représentation de la *possibilité nuancées* d'un évènement



N(A) = 1 - Π(nonA) = min(1 - π(u) | u ∉ A)

N(A) : degré d'inclusion : 'JEUNE' ⊆ A
On calcule à quel point A se déduit de l'information "JEUNE"

Fig. III.5 (D) : Représentation de la *nécessité nuancées* d'un évènement

Figure III.5 : Formes de représentation possibiliste de l'information imprécise [86]

III.3.4 La logique évidentielle (théorie de croyance)

III.3.4.1 Notions de base

Les fonctions de croyance de Dempster (1967) [73] et Shafer (1976) [74] englobent la théorie des possibilités et des probabilités dans un cadre discret. Par ailleurs, la question du conflit demeure essentielle dans la théorie des fonctions de croyance, ce qui rend l'étude du conflit particulièrement pertinente dans ce contexte.

L'objectif de cette théorie, vue comme une généralisation de la théorie des probabilités, est de combiner des preuves distinctes afin de déterminer la probabilité d'un événement. Cette approche offre des avantages par rapport à la théorie bayésienne, car elle permet de :

- Considérer en plus des hypothèses simples, les unions (combinaison, fusion) des hypothèses simples.
- Modéliser le conflit entre les informations issues de différentes sources et de les combiner afin de prendre une meilleure décision.
- Modéliser l'inconsistance et l'incertitude

La logique évidentielle, basée sur le théorème de Dempster-Shafer, est basée sur les notions suivantes [87] :

- **Cadre de discernement :** un cadre de discernement désigne un ensemble qui regroupe toutes les hypothèses possibles relatives au problème analysé. Appelons cet ensemble Ω et supposons qu'il se compose de N hypothèses distinctes :

$$\Omega = \{H_1, H_2, ..., H_N\}$$

Une seule hypothèse de cet ensemble est jugée vraie. L'ensemble des parties (ou power set en anglais) du cadre de discernement est celui qui regroupe toutes les combinaisons possibles des hypothèses. L'ensemble des parties se définit comme suit :

$$2^{\Omega} = \{A|A \subseteq \Omega\} = \{\{H_1\}, \{H_2\}, \{H_1, H_2\}, \{H_1, H_3\} ..., \{H_1, H_N\}, ..., \Omega\}$$

A peut-être une hypothèse simple ou un ensemble d'hypothèses ou même un ensemble vide.

- **Fonctions de masse :** en général, une fonction de masse, souvent appelée masse de croyance ou BPA (Basic Probability Assignment), est simplement notée m. Cette masse est déterminée pour chaque élément de l'ensemble 2^{Ω} et se définit comme suit : $\forall A \in 2^{\Omega}$, $m(A) \in [0, 1]$;

$$\sum_{A \in 2^{\Omega}} m(A) = 1 \qquad \text{(III.2)}$$

La masse m(A) d'un élément A spécifique de l'ensemble des parties représente la proportion de toutes les preuves disponibles qui soutiennent que l'état actuel est A, et non un autre état ou un sous-état de A. Ainsi, la valeur de m(A) reflète uniquement la confiance accordée à l'état A sans attribuer de crédit aux sous-ensembles (hypothèses simples ou composées) de A, chacun ayant, par définition, sa propre masse.

Un élément ayant une masse non nulle est qualifié d'élément focal.

III.3.4.2 Mesures de croyance et de plausibilité

La croyance peut être attribuée à d'autres fonctions qui sont associées à la fonction de masse m :

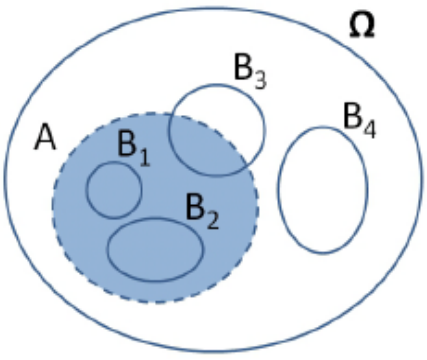
- **Fonction de croyance (crédibilité) :** la crédibilité (belief), notée bel, évalue la croyance totale pouvant être associée à un élément spécifique. Elle se définit par :

$$\forall A \in 2^{\Omega}, Bel(A) = \sum_{B \in A, B \subseteq \Omega} m(B) \qquad \text{(III.3)}$$

- **Fonction de plausibilité** : la plausibilité (plausibility), notée pl, évalue la croyance maximale qui pourrait être associée à un élément spécifique. Elle se définit par :

$$\forall A \in 2^{\Omega}, Pl(A) = \sum_{B \cap A \neq \emptyset} m(B) \qquad \text{(III.4)}$$

De plus, à partir de la valeur de la masse d'un état, il est possible de définir un intervalle de probabilité. Cet intervalle renferme la valeur exacte de la probabilité de l'état et est délimité par deux mesures, nommées croyance et plausibilité : Bel(A) ≤ Prob(A) ≤ Pl(A)



Bel(A) = m(B_1) + m(B_2)

Pl(A) = m(B_1) + m(B_2) + m(B_3)

$$bel(A) = \sum_{\emptyset \neq B \subseteq A} m(B), \quad \forall A \subseteq \Omega$$

Bel(A) : représente la part *totale* de croyance soutenant A

$$pl(A) = \sum_{B \cap A \neq \emptyset} m(B), \quad \forall A \subseteq \Omega$$

Pl(A) : représente la part *maximale* de croyance qui pourrait soutenir A

Figure III.6 : Fonction de croyance et fonction de plausibilité [88]

III.4 Comparatif entre les déférentes logiques de modélisation

Il n'existe pas de logique de modélisation universelle, chaque approche est conçue pour traiter un type spécifique d'incertitude. La logique binaire s'inscrit dans une vision strictement déterministe où l'information est soit vraie, soit fausse, tandis que la logique probabiliste s'appuie sur des données statistiques afin de représenter l'aléa. À l'inverse, la logique floue permet de modéliser des concepts imprécis et graduels, proches du raisonnement humain. La logique possibiliste se révèle particulièrement pertinente lorsque les connaissances sont incomplètes, alors que la logique évidentielle offre une distinction explicite entre ignorance et croyance grâce à la combinaison de plusieurs sources d'information. Par conséquent, le choix d'une logique de modélisation dépend étroitement du contexte d'application, de la nature des données disponibles et des objectifs poursuivis. Le tableau suivant présente un comparatif, non exhaustif, entres les logiques de modélisation les plus utilisées :

Logique	Incertitude	Contraintes $\mu(x)$	Usage & Sémantique
Binaire	Aucune	$\mu \in \{0, 1\}$	- Classes strictement séparées - Déterminisme pur
Probabiliste	Aléatoire	$\Sigma \mu = 1$	- Mesure de fréquence d'occurrence - Fondé sur la théorie de la mesure
Floue	Lexicale	$\mu \in [0, 1]$	- Classes chevauchées - Traite l'imprécision du langage, ex : *chaud*
Possibiliste	Épistémique	$N(A) \leq P(A) \leq \Pi(A)$	- Distingue le *possible* Π du *nécessaire* N - Idéal pour les données incomplètes
Évidentielle	Subjective	$Bel(A) \leq Pl(A)$	- Fusion de sources d'opinions - Gère l'ignorance totale et le conflit

Table III.1 : Comparatif entre les logiques de modélisation de l'incertitude

Pour rendre ce comparatif plus exhaustif, il est pertinent d'y intégrer des logiques avancées capables de traiter des formes plus complexes d'incertitude et d'imprécision, comme :

- **La logique des Rough Sets (ensembles approximatifs)** : particulièrement adaptée au traitement de l'indiscernabilité entre données, en s'appuyant sur les notions d'approximation inférieure et supérieure sans nécessiter d'information probabiliste préalable.
- **La logique intuitionniste floue** : étend la logique floue classique en introduisant, en plus du degré d'appartenance, un degré de non-appartenance indépendant, ce qui permet de modéliser explicitement l'hésitation et l'incertitude résiduelle.
- **La logique neutrosophique** généralise ces approches en intégrant une composante d'indétermination I distincte aux degrés de vérité T et de fausseté F, offrant ainsi un cadre adéquat pour la modélisation des Big Data hétérogènes ou corrompues [89].

Partie 2 : Adaptation de l'algorithme Apriori aux logiques de l'incertain

III.5 Les règles d'association floues

III.5.1 Présentation générale

Comme on a cité auparavant, l'algorithme Apriori d'Agrawal [54] et ses variantes ont été conçus pour des bases de données booléennes, (on met la valeur « 1 » dans la base des transactions si le produit est acheté, sinon on met la valeur « 0 »). Dans ce cas, la logique propositionnelle simple est suffisante pour exprimer les RA. Mais nous n'avons pas pris en considération des critères comme par exemple la quantité des produits achetés. Normalement, un produit acheté pour une quantité donnée n'est pas considéré comme le même produit acheté avec une quantité différente, aussi les données sont représentées en logique binaire en négligeant tous aspect incertain ou imprécis ou nuancé des données analysées. Donc, les algorithmes classiques d'extraction des RA s'avèrent alors inadaptés pour traiter ce type d'informations. Pour cela, et pour pallier ce problème, les RA floues sont apparues. C'est une modélisation des RA basées sur la théorie des sous-ensembles flous proposée par Lotfi Zadeh [64], permettant de raisonner sur des attributs quantitatifs ou ayant un aspect nuancé ou incertain. Cette adaptation de l'Apriori sera présentée en ce que suit.

III.5.2 Concepts et définitions de base

En se basant sur la théorie des SEF, les nouvelles définitions d'un *item* et *itemset* flous, du *support* d'un itemset flou et la *confiance* d'une *RA* floue sont présentés ci-dessous.

III.5.2.1 Item flou

Un item flou est un couple [item, SEF], où on associer un item avec l'un des Sous-Ensembles Flous de quantité associes. Cet item flou noté [x, A] correspond à l'item x, associe au SEF A, défini sur l'univers du discours des valeurs de x.

Exemple : [température, froide] est un item flou ou *froide* est un SEF défini par sa Fonc_App

III.5.2.2 Itemset flou

Un itemset flou, noté par (X, A), est un ensemble d'items flous, où $X = \{x_1, x_2, \ldots, x_p\}$ est un ensemble de « p » items et $A = \{a_1, a_2, \ldots, a_\ell\}$ est un ensemble de « ℓ » SEF associes à ces items. Par exemple ([température, froide] [humidité, haute]) est un itemset flou avec deux items flous [température, froide] et [humidité, haute]. Un itemset flou (X, A) n'est pas valide s'il contient au moins deux items flous du même attribut. Exemple : ([température, froide] [température, chaude]) n'est pas un itemset valide parce qu'il contient deux items flous du même attribut. Un K-itemset flou est l'ensemble d'itemsets flous contenant k items flous.

III.5.2.3 Redéfinition de l'union et de la cardinalité des ensembles

L'adaptation de l'extraction des RA à la logique floue nécessite une redéfinition des opérations de base de la théorie classique des ensembles, afin de traiter les *ensembles chevauchés*. Ces redéfinitions touchent notamment les notions de l'*union* d'items (pour créer un itemsets flous et en déduire son *Deg_App*) et de la *cardinalité* (pour compter la fréquence d'apparition des itemsets dans toutes les transactions et en déduire leur support). Ainsi, le calcul du support des itemsets flous (et donc la confiance de la RA floue extraite) est lié étroitement à la redéfinition des deux opérations de base suivantes [90] :

- ***Union des items flous :*** pour calculer le *Deg_App* des itemsets flous, nous utilisons la notion de normes triangulaires T, notées :
  - T-norme (⊤) : comme généralisation de l'opérateur d'intersection ;
  - T-conorme (⊥) : comme généralisation de l'opérateur d'union.
- ***Cardinalité de l'itemset flou :*** le support d'un itemset flou est le pourcentage de transactions supportant (contenant) l'itemset flou (X,A) sur le nombre total de transactions. Le calcul du support de l'itemset flou repose sur une redéfinition de la cardinalité de l'itemset flou, qui correspond à sa fréquence d'apparition dans les lignes de la table des transactions. La cardinalité de l'itemset flou (X,A) est liée à son *Deg_App*.

III.5.2.4 Le degré d'appartenance d'un itemset flou

Chaque item flou (x_i, a_j) est représenté dans une transaction t par son degré d'appartenance *Deg_App*, noté $\mu(a_j)$, sachant que : $(\mu(a_j)t[x_i]) \in [0,1]$. Cela signifie que le *Deg_App* de l'item x_i au SEF a_j dans la ligne (la transaction) t est : $\mu(a_j)$. Un item flou, ou un itemset flous, est dit fréquent si son *Deg_App* $\geq \Omega$ (seuil minimal défini par un expert). Trois approches ont été proposées dans la littérature pour calculer le *Deg_App* d'un itemset flou (X, A) dans une transaction partir des *Deg_App* de ses items :

- Delgado et al. [91] ont utilisé l'opérateur "Min" comme une t-norm (⊤), c'est le minimum entre les *Deg_App* des items flous qui composent l'itemset flou (X, A) dans une transaction.
- Fiot et al. [93] ont utilisé l'opérateur "Max" comme t-conorme (⊥), c'est le maximum entre les *Deg_App* des items flous composant l'itemset flou (X,A) dans une transaction.
- Kuok et al. [67] ont préféré d'utiliser l'opérateur "Produit" Π des *Deg_App* des items flous qui composent l'itemset flou (X, A) dans une transaction.

Pour expliquer ces trois approches, le tableau suivant (Tab. III.2) donne un exemple de calcul des *Deg_App* de chaque item x_i aux différents SEF pour chaque ligne (transaction) t_j.

Transaction	L'item x_1			L'item x_2		*Deg_App* de l'itemset flou ([x_1, $a_{1.3}$] [x_2, $a_{2.2}$])		
	$a_{1.1}$	$a_{1.2}$	$a_{1.3}$	$a_{2.1}$	$a_{2.2}$	t-norm (⊤)	t-conorme (⊥)	Produit (Π)
t_1	0.1	0.1	0.8	0.1	0.9	0.8	0.9	0.72
t_2	0.1	0.4	0.5	0.5	0.5	0.5	0.5	0.25
t_3	0.1	0.7	0.2	0.7	0.3	0.2	0.3	0.06
t_4	0.6	0.4	0	0.75	0.25	0	0.25	0

Table III.2 : Exemple de calcul du Deg_App d'un itemset flou via les trois approches

Le tableau suivant présente les normes triangulaires les plus utilisées :

t-norme	t-conorme	Nom
Min(x, y)	Max(x, y)	Zadeh
$x.y$	x + y – x.y	Probabliste
Max($x + y - 1$, 0)	Min($x + y$, 1)	LukasieWicz
$x.y$	$x + y - x.\ y - (1 - \gamma).\ x.\ y$	Hamacher ($\gamma > 0$)
$\begin{cases} x\ si\ y = 1 \\ y\ si\ x = 1 \\ 0\ sinon \end{cases}$	$\begin{cases} x\ si\ y = 0 \\ y\ si\ x = 0 \\ 0\ sinon \end{cases}$	Weber

Table III.3 : Les t-normes et t-conormes les plus utilisés [98]

III.5.2.5 Le support d'un itemset flou

Pour calculer le support flou d'un itemset flou, quatre approches basées sur la cardinalité d'un itemset flou ont été proposées [93] :

- Comptage_binaire : on compte toute transaction contenant l'itemset (X, A).
- Comptage_seuillé : on comptabilise toutes les transactions pour lesquelles le degré d'un itemset (X, A) est supérieur à un seuil fixe (α-coupe).
- Σ Comptage : sommer les degrés d'un itemset (X, A) de chaque transaction.
- Σ Comptage_seuillé : sommer les degrés d'un itemset (X, A) au-delà d'un seuil minimal fixe (α-coupe) de chaque transaction.

Pour l'exemple précèdent, les différents calculs de cardinalité de l'itemset (X, A) = ([x_1, $a_{1.3}$] [x_2, $a_{2.2}$]) en utilisant un t-norm (⊤) sont montrés dans le tableau suivant (Tab. III.4) :

Transactions	Degré de l'itemset flou ([x_1, $a_{1.3}$] [x_2, $a_{2.2}$])			
t_1	1	1	0.8	0.8
t_2	1	1	0.5	0.5
t_3	1	0	0.2	0.2
t_4	0	0	0	0
Type de comptage	Binaire	Comptage_seuillé (α-coupe = 0.4)	Σ Comptage	Σ Comptage_seuillé (α-coupe = 0.4)
	3	2	1.5	1.3

Table III.4 : Diffèrent type de calcul de la cardinalité floue

A partir de ces différentes méthodes de calcul du Deg_App d'un itemset flou, on peut trouver plusieurs supports. Par exemple, si on choisit l'opérateur t-norm (⊤) avec un Σ Comptage_seuillé, le support d'un itemset flou (X, A) est calculé comme suit :

$$Supp(X, A) = \frac{\sum_{t_i \in T} \top\left(\alpha_A(t_i[X])\right)}{|T|} \qquad \text{(III.5)}$$

Où |T| est le nombre de transactions dans la base de transactions T, et α représente la fonction d'appartenance souillée de l'équation suivante :

(III.6)

$$\alpha_A(t_i[X]) = \begin{cases} \mu_A(t_i[X]), & si\ \mu_A(t_i[X]) > \alpha - coupe \\ 0, sinon \end{cases}$$

Où, $\mu_A(t_i[X])$ est un vecteur de p valeurs floues, p indique le nombre des items de l'itemset X, et chaque valeur floue représente le *Deg_App* de chaque item $x_{i=1\ldots p} \in X$ à ces SEF associes a_i (i = 1… ℓ) ∈ A, et α-coupe est un seuil minimum d'appartenance définie par l'utilisateur.

III.5.2.6 Les règles d'association floues

Une RA floue a la forme : Si "X est A", alors "Y est B". Où la partie "X est A" est la condition de la règle et la partie "Y est B" sa conclusion. Cette règle s'écrit : (X,A) → (Y,B), où X ={x_1, x_2, . . . , x_p} et Y ={y_1, y_2, . . . , y_k} sont deux itemsets flous disjoints, tel que A = {a_1, a_2, . . . , a_ℓ} et B = {b_1, b_2, . . . , b_m} sont des SEF associés aux itemsets flous X et Y.

III.5.2.7 La confiance d'une règle d'association floue

Si on choisit l'opérateur t-norm (⊤) avec un Σ Comptage_seuillé, la confiance d'une règle d'association floue (X, A) → (Y, B) est calculée comme suit :

$$Conf(X,A) \rightarrow (Y,B) = \frac{\sum_{t_i \in T}[\alpha_A(t_i[X])] \top [\alpha_B(t_i[Y])]}{\sum_{t_i \in T} \alpha_A(t_i[X])} \qquad \text{(III.7)}$$

Où $\alpha_A(t_i[X])$ et $\alpha_B(t_i[Y])$ ont la même définition décrite dans l'équation (III.6)

III.5.3 Extraction de règles d'association floues

Tout d'abord, chaque item (attribut) est partitionné en plusieurs SEF afin de construire les items flous en transformant les données quantitatives de la table des transactions en *Deg_App* à ces SEF. Les SEF et la forme des fonctions d'appartenance sont définis par des experts du domaine ; mais, des fois cette expertise n'est pas toujours présente, soit à cause du manque des experts dans certain domaine ou bien à cause de la complexité du problème. Pour cela des algorithmes sont utilisés afin d'obtenir un partitionnement de chaque attribut en SEF tel que l'algorithme de C-Moyenne Flous (FCM) [94]. Une fois les SEF sont déterminés et les *Deg_App* sont calculés, cette table sera utilisée afin d'extraire les RA floues. Pour découvrir ces règles une variété d'approches ont été développées.

Plusieurs algorithmes classiques d'extraction des RA ont été fuzzifiés en intégrant les concepts de la théorie des SEF. Par exemple, l'algorithme de Pierrard et al. [95] repose sur l'algorithme Close [96], l'algorithme de Hong et al. [92] est basé sur l'algorithme Apriori [54], aussi Hong et al [97] ont proposés l'algorithme Apriori-TID flou fondé sur l'algorithme Apriori-TID [54]. Ces différents algorithmes développés génèrent tous des RA floues en se basant sur le même principe de redéfinition du calcul du support et de la confiance. Par exemple, Hong et al. [92] ont proposés d'utiliser l'opérateur "Min" comme une t-norme avec la cardinalité "Σ Comptage". Une autre méthode a été proposée par Kuok et al. [67], ils ont utilisé deux facteurs pour valider les RA floues, un facteur de signification et un facteur de certitude. Une règle est valide si est seulement si ses deux facteurs sont supérieurs à des seuils définis par

l'utilisateur, le premier facteur est basé sur la notion du support, il a utilisé l'opérateur produit (Π) avec un Σ Comptage_seuillé pour le calculer, le facteur de certitude est le ratio du facteur de signification de l'itemset constituant la règle sur le facteur de signification de la condition.

III.5.4 Exemple d'extraction de règles d'association floues

- **Base de données :** Soit la base des transactions (voir Tab. III.5) qui comporte la quantité achetée pour chaque article (item).

Transactions	Item 1	Item 2	Item 3	Item 4
t_1	1	1	5	4
t_2	1	2	4	3
t_3	3	2	4	1
t_4	2	2	5	2
t_5	1	3	3	2
t_6	2	3	4	1

Table III.5 : Base de transactions quantitatives utilisée

- **Représentation des fonctions d'appartenances :** Tous d'abord il faut convertir la base de données quantitative en base de données de degrés d'appartenance. Les partitions des SEF de chaque item quantitatif et les fonctions d'appartenance pour chacun de ces SEF sont illustrées par la figure III.7). Par exemple, l'item2 est découpe en 3 parties : peu (p), moyen (m), beaucoup (b), l'achat d'un seul item2 est considéré comme appartenant au SEF "peu", 2 comme quantité "moyenne", 3 quantités de l'item2 correspondent à la fois une à une quantité "moyenne" et "beaucoup", 4 quantités de l'item2 et plus est considéré comme "beaucoup". A partir des fonctions d'appartenance ci-dessus, on définit la base transactionnelle floue (Tab III.6), qui donne les *Deg_App* de chaque item à chaque SEF.

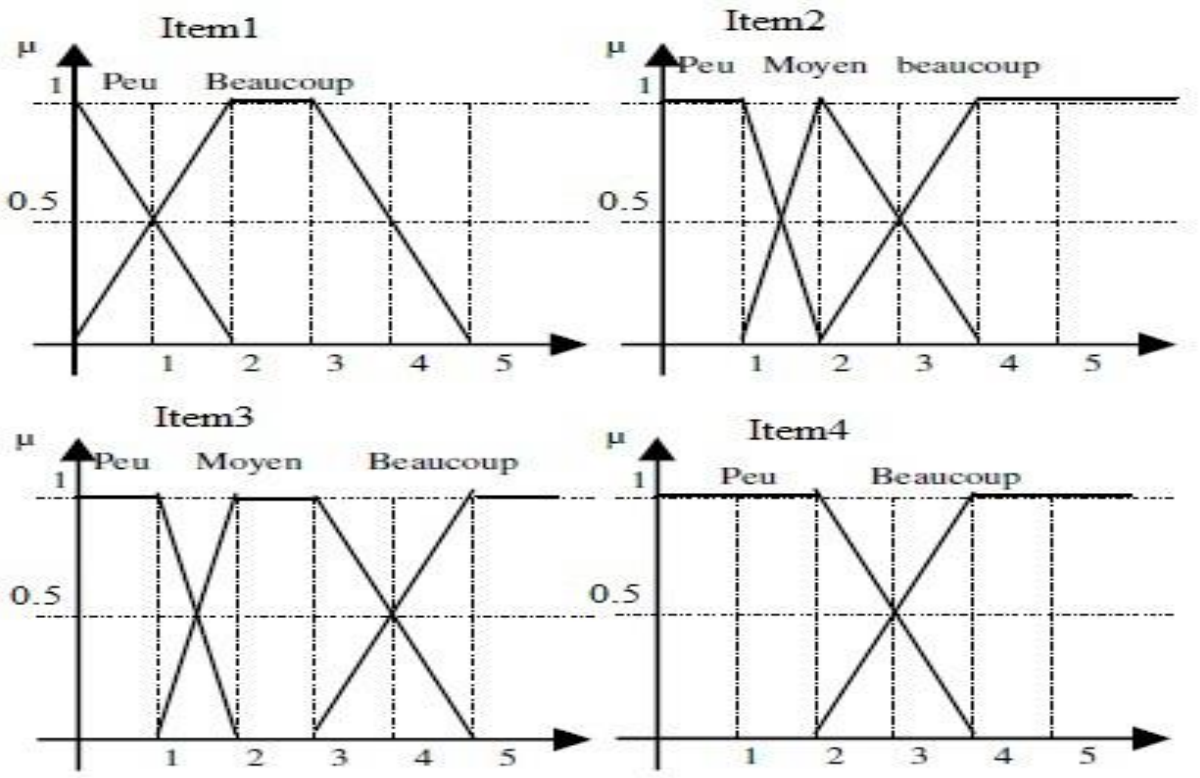


Figure III.7. Partitionnement flou des attributs quantitatifs

La table des transactions floue résultante est la suivante :

Transactions	Item 1		Item 2			Item 3			Item 4	
	P	B	P	M	B	P	M	B	P	B
t_1	0.5	0.5	1	0	0	0	0	1	0	1
t_2	0.5	0.5	0	1	0	0	0.5	0.5	0.5	0.5
t_3	0	1	0	1	0	0	0.5	0.5	1	0
t_4	0	1	0	1	0	0	0	1	1	0
t_5	0.5	0.5	0	0.5	0.5	0	1	0	1	0
t_6	0	1	0	0.5	0.5	0	0.5	0.5	1	0

Table III.6 : Table des transactions floues

- **Recherche des items fréquents :** Les items flous correspondant à la base exemple sont les suivants : [Item1, peu], [Item1, beaucoup], [Item2, peu], [Item2, moyen], [Item2, beaucoup], [Item3, peu], [Item3, moyen], [Item3, beaucoup], [Item4, peu], [Item4, beaucoup].
  Si on fixe les seuils α-coupe = 0,5 et un support minimal *Min_Sup* = 0.45, les items flous fréquents sont (voir tableau III.7) : [Item1, beaucoup], [Item2, moyen], [Item3, beaucoup] et [Item4, peu].

Item	SEF	Support flou
Item 1	Peu	0.25
	Beaucoup	0.75
Item 2	Peu	0.17
	Moyen	0.67
	Beaucoup	0.17
Item 3	Peu	0
	Moyen	0.42
	Beaucoup	0.58
Item 4	Peu	0.75
	Beaucoup	0.25

Table III.7 : Support flou des items

- **Calcule du support et confiance d'une règle d'association floue :** par exemple, le support et la confiance en utilisant l'équation III.5 et III.7 pour la règle : [Item2, moyen] → [Item1, beaucoup]

$$Support = [min(0,0.5)+min(1,0.5)+min(1,1)+min(1,1)+min(0.5,0.5)+min(0.5,1)] / 6$$
$$= [0.5+1+1+0.5+0.5] / 6 = 0.583$$

$$Confiance = [min(0, 0.5)+min(1, 0.5)+min(1, 1)+min(1, 1)+min(0.5, 0.5)+min(0.5, 1)] /$$
$$0+1+1+1+0.5+0.5 = 3.5 / 4 = 0.875$$

Ce qui signifie que : si achète_moyen de l'item2 alors achète_beaucoup de l'item1 dans 87.5% des cas. Moyen de l'item2 et beaucoup de l'item1 sont achetés ensemble dans 58.3% des opérations d'achats.

III.6 Etat de l'art : Travaux sur les Exp_Gen basés sur les RA

La recherche des règles d'association pertinentes est l'un des principaux algorithmes de Machine Learning pour analyser les données d'expression génétique. Tous les algorithmes conventionnels d'extraction des AR consomment du temps CPU et génèrent un très grand nombre de RA parfois inintéressantes, et le défi sera de déterminer quelles règles sont les plus utiles, via des mesures d'intérêt appropriées. Dans cette perspective, on peut répartir les travaux dans la littérature en deux axes : sélection des RA suivant des mesures d'intérêts subjectifs ou objectifs.

III.6.1 Sélection des RA pertinentes basée sur des mesures d'intérêt subjectifs

La sélection des RA les plus intéressantes en se basant sur des *mesures d'intérêt subjectif* (des RA biologiquement interprétables) était le sujet de recherche des travaux suivants :

Pour extraire des RA suivant l'intérêt fonctionnel des gènes qui les composent, [99] propose un ARM dynamique pour déterminer si l'induction / répression des gènes correspond aux variations phénotypiques, y compris les régulations cellulaires, les diagnostics cliniques et le développement de médicaments. [100] proposent un AR générique pour extraire des groupes de gènes biologiquement significatifs pouvant exhiber en corrélations négatives.

Pour développer une approche de médecine personnalisée, [101] propose une méthode de Deep Learning associée à ARM pour prédire la réponse aux médicaments du cancer, cela via l'extraction des signatures spécifiques au cancer sous la forme de règles facilement interprétables, utilisées pour prédire les réponses pharmacologiques à un grand nombre de médicaments anticancéreux. [102] a réalisé une méta-analyse des données de puces à ADN (microarray) liées au cancer, qui nous a permis d'identifier un ensemble de dix lncRNA altérés simultanément dans différentes bases de données sur les tumeurs cérébrales. Aussi, [103] ont appliqué le t-test statistique pour sélectionner les gènes significatifs, le k-Means pour discrétiser les données Micro-array d'expression génétique du cancer, Boolean ARM pour générer les intervalles d'expression génique fréquents et enfin, les AR ont été découvertes pour fournir la décision significative pour le diagnostic du cancer. Ensuite, [104] ont amélioré leur travail pour pouvoir exposer la relation entre les données normales et anormales d'expression génique, et permettent de prendre des décisions pour un traitement génétique ciblé.

III.6.2 Sélection des RA pertinentes basée sur des mesures d'intérêt objectifs

La sélection des RA basée sur des *mesures d'intérêt objectifs* (intérêt purement computationnel) était détaillée dans les travaux de recherche suivants :

[105] proposent deux formes de ARM à partir d'images ISH : des RA spatiales entre les zones d'expression génique et des RA entre les gènes qui expriment. Leur algorithme d'ARM

est basé sur une distance de Jaccard. Aussi, en plus du support et de la confiance, [106] propose d'utiliser l'entropie et la variance comme mesures d'intérêt pour extraire des AR spatiales d'expression génétique.

En se basant sur les algorithmes bio-inspirés, [107] ont utilisé les Ant-Colony Algorithms pour générer le nombre optimal de clusters pour la segmentation des données d'expression génique ; ces Ant-ARM sont proposés pour extraire des règles de classification associative. Aussi, [108] propose une nouvelle approche de recherche RA basée sur l'algorithme d'optimisation de recherche des pingouins, qui intègre une mesure d'évaluation de la quantité de chevauchement entre les RA générées pour choisir un nombre limité de règles de haute qualité. [109] proposent des RA pour les données d'Exp_Gen en utilisant une approche basée sur les réseaux de neurones pour décider laquelle des règles est fortement exprimée.

Les méthodes de clustering ont également été utilisées pour sélectionner les AR à haut intérêt. [110] a converti les données d'expression génique en intervalles d'expression génique en utilisant une technique de discrétisation. L'algorithme de clustering agglomératif à liaison unique est utilisé pour regrouper les AR dérivées sur la base d'une matrice de similarité calculée à partir d'une distance euclidienne. En outre, [111] propose une features (gène) sélection basée sur les mesures d'impureté statistique (indice de Gini, minorité maximale et Twoing rule), puis calcule la similitude entre les AR découvertes basée sur une distance de Jaccard pondérée et sur la mesure de similarité du cosinus vectoriel. Enfin, un regroupement des RA via un clustering hiérarchique a été proposé.

D'après l'état de l'art, on remarque que peu d'efforts ont été fournis concernant la modélisation floue des données d'expression génétique qui sont bruitées de nature et à aspect imprécis, et aucune modélisation floue n'a été proposée pour analyser les données ISH. De plus, aucun travail publié (à notre connaissance) n'a traité ces données d'Exp_Gen en les formulant via une logique possibiliste, qui est une vision plus large de la logique floue, pour les analyser via n'importe quelle méthode computationnelle, notamment les Règles d'Association.

III.7 Conclusion

Pour préparer le cadre théorique de l'approche proposée, ce chapitre a présenté dans sa première partie les notions de base des différentes logiques de l'incertain, utilisées pour la modélisation de l'information imparfaite intégrée dans les datasets à analyser. Dans la deuxième partie, on a expliqué les adaptations nécessaires de l'algorithme Apriori pour le rendre en adéquation avec les logiques de l'incertain, dans le but d'extraire les relations cachées dans les données d'expression génétique qui sont de nature nuancée et donc loin d'être en logique binaire.

Dans le chapitre suivant, on présentera les détails de l'approche proposée et les pertinences des connaissances extraites via notre modélisation.

Chapitre 4

Chapitre IV : Approche proposée ET Interprétation biologique des résultats

IV.1 Introduction

Ce chapitre présente l'essentiel du travail de recherche publié dans le journal *Engineering, Technology & Applied Science Research*, Vol. 15, No. 4, 2025, 25304-25312. La première partie de ce chapitre présente les détails de l'approche proposée, commençant par un cadre théorique, un pseudo algorithme, et l'avantage de la contribution fournie par cette approche. La deuxième partie de ce chapitre présente les étapes d'implémentation de notre approche. On expliquera les différentes étapes de prétraitement qui génèrent les tables de transaction suivant les deux modélisations proposées ; ces tables seront traitées par l'algorithme Apriori préalablement adapté à ses deux logiques. Une représentation graphique et tabulaire des résultats (en primal et en dual) sera fournie. On finira par une analyse et validation des résultats des deux approches (floue/possibiliste), ainsi qu'une interprétation biologique des connaissances extraites via la plateforme String-DB.org et une discussion et évaluation de la pertinence de ces résultats, pour décider quelle est la logique de modélisation la plus adéquate à la réalité biologique des images ISH étudiées.

Partie 1 : Approche Proposée

IV.2 Justification de l'approche proposée

IV.2.1 Présentation générale

L'augmentation phénoménale du volume de données spatio-temporelles a engendré divers problèmes de modélisation et d'interprétation de ces données complexes, qui dissimulent une grande quantité de connaissances [112].

Aussi, analyser des données biologiques hétérogènes et vitales, en pleine expansion et souvent imprécises et/ou incomplètes, et interpréter les connaissances extraites reste toujours un défi. Les images ISH d'expression génétique durant les stades de développement embryonnaire de l'espèce modèle "Edinburgh Mouse", représentent un atlas de données biologiques spatio-temporelles ; leur étude est essentielle pour la pré-détection de maladies d'origine génétique durant les phases préliminaires de développement des espèces vivantes.

Les règles d'association permettent d'extraire des relations cachées entre un ensemble d'attributs initialement de type binaire. L'extraction des RA est l'un des algorithmes d'apprentissage automatique les plus utilisés pour analyser les données d'Exp_Gen, notamment pour sélectionner les RA les plus intéressantes en fonction de mesures d'intérêt objectif (intérêt purement computationnel) [105] ou subjectif (RA biologiquement interprétable) [100] [104].

Mais, dans le monde réel, on traite généralement des connaissances ambiguës sur toute situation, soit parce qu'on doute de leur véracité (c'est l'*incertitude*), soit parce qu'on a des difficultés à les énoncer clairement (c'est l'*imprécision*) [89]. Par conséquent, la représentation binaire des données réelles entraîne généralement une perte importante de précision dans leur quantification et leur sémantique. Dans la littérature, des adaptations de l'algorithme Apriori ont été proposées pour se rapprocher de plus en plus du raisonnement humain (qui est généralement nuancé et loin d'être binaire) et de la réalité des données analysées, grâce à la prise en charge de données de nature incertaine et/ou imprécise, via une modélisation floue [113], et la modélisation possibiliste [72]. Ces RA adaptées permettent d'extraire des connaissances vitales.

IV.2.2 Critique des travaux antecedants

Trois principales technologies de séquençage des données d'Exp_Gen sont étudiées dans la littérature :

- Les puces d'ADN (DNA microarray data) : la majorité des travaux se concentrent sur la sélection de caractéristiques intéressantes pour une classification efficace des gènes [115], ou

sur la recherche de distances et de mesures de similarité adéquates pour le clustering et la découverte de motifs cachés [22].

- Séquençage d'ARN (RNAseq) : travaux axés sur la sélection de caractéristiques pour la classification [28] ou le clustering [116].

- Images ISH (séquences d'ADN ou d'ARN d'images à haute résolution) étudiées essentiellement pour : la détection de gènes co-exprimant [34] et leur clustering pour extraire des motifs cachés [35].

Dans notre étude, nous avons analysé les images ISH d'Exp_Gen. Dans des travaux antérieurs [105] [100] [104], seule une représentation binaire a été utilisée pour la modélisation, ce qui peut nuire à la qualité des connaissances extraites, car les données étudiées étaient considérées comme précises et parfaites. Mais en réalité, ces données peuvent cacher des incertitudes (il n'est pas strictement certain qu'un gène soit toujours exprimé dans telle zone de pixels, telle phase et tel niveau d'expression) et des imprécisions (les frontières des zones d'Exp_Gen ne sont pas clairement définies). Par conséquent, ignorer une éventuelle incertitude ou imprécision dans ces données peut diminuer la pertinence des résultats des approches déjà proposées dans la littérature.

Ainsi, nous constatons que peu d'efforts ont été fournis concernant l'extraction de RA floues à partir de données d'Exp_Gen [113] [117], et qu'aucune modélisation RA floue n'a été proposée pour l'analyse des images ISH d'Exp_Gen, qui sont par nature nuancées, et nous ne pouvons pas négliger la possibilité d'incertitude et d'imprécision dans les limites des zones et l'intensité des couleurs d'expression génétique (voir Figure I.7 du chapitre I). De plus, aucune publication (à notre connaissance) n'a étudié ce type de données d'Exp_Gen en les formulant via une logique possibiliste, qui est une vision plus large de la logique floue, pour les analyser via une méthode computationnelle, y compris les RA.

IV.2.3 Proposition

Dans cette perspective, notre travail vise à proposer une étude fidèle à la réalité biologique des images ISH d'Exp_Gen, résultant du suivi des stades de développement de l'espèce modèle « Souris d'Edinbourg » [10]. Nous nous intéressons à l'adaptation de l'algorithme Apriori aux théories de l'incertitude pour modéliser les données analysées afin d'extraire plusieurs types de RA utiles, selon les deux modélisations proposées (floue et possibiliste). De plus, la nature nuancée des forces d'Exp_Gen (forte, modérée et non détectée) sera prise en considération pour extraire les caractéristiques les plus intéressantes qui représentent la réalité de ces données biologiques spatiotemporelles.

Après une série d'opérations de prétraitement et d'extraction de caractéristiques visant à réduire la dimensionnalité des images étudiées (et donc la complexité algorithmique) tout en préservant leur variabilité, nous cherchons à extraire : les relations entre les Exp_Gen en détectant les ensembles de gènes co-exprimants, les régions dont l'expression est synchronisée et les relations temporelles entre les gènes exprimés à différents stades du développement

embryonnaire. Les RA extraites permettent de découvrir diverses connaissances biologiques cachées dans ces images ISH.

L'adaptation proposée de l'algorithme Apriori (initialement conçu pour les données binaires) repose sur la redéfinition des notions de *cardinalité* et d'*union* entre itemsets. De plus, nous utilisons une méthode basée sur des *normes triangulaires* pour calculer le support des itemsets et la confiance de la RA extraite, afin de les adapter aux logiques floues et possibilistes.

Enfin, une analyse et une comparaison de la qualité des règles extraites des deux modèles flous et possibilistes seront proposées afin de déterminer la modélisation la plus adéquate. De plus, une analyse fonctionnelle des gènes formant les itemsets des RA extraites confirme qu'une grande partie des co-expressions détectées est conforme aux principes biologiques, ce qui confirme que ces gènes collaborent réellement lors de la formation de différents systèmes biologiques.

IV.3 Adaptation proposée de l'Apriori aux théories de l'incertain

En logique binaire (base de l'algorithme Apriori), le degré de véracité d'une proposition ne peut prendre que l'une des deux valeurs {1, 0} (vrai ou faux), ce qui est inadapté à la modélisation (et au traitement) de la majorité des données du monde réel, notamment les données d'expression génétique, qui sont loin d'être binaires. Par la suite, [64] ont développé un nouveau cadre théorique pour la modélisation de l'incertitude, la théorie des Sous-Ensembles Flous (SEF), qui permettait de manipuler une véracité nuancée entre complètement vrai et complètement faux, étendue ensuite à la logique possibiliste [71]. En ce qui suit, nous présenterons les concepts de la théorie des SEF et de la logique possibiliste à adapter pour rendre l'algorithme Apriori adéquat aux données biologiques nuancées étudiées dans ce travail.

IV.3.1 Concepts de la théorie des SEF adaptés à notre approche

La théorie SEF intègre le concept de vérité partielle (ou véracité graduelle). Pour une modélisation en logique floue nous devons définir [64] :

Univers de Discours (Univ-Dis) : termes utilisés dans un domaine d'étude. Dans notre approche ce sont les forces d'expression génétique : Forte (couleur rouge), Moyenne (jaune) et Non Détectée (gris). Chacun de ces termes (couleur) formera un SEF et est représenté par :

- ✓ Dans notre approche, l'Univ_Dis X est l'intensité de couleur d'expression génétique (de 0 à 255), et x est un pixel dans l'image d'Exp_Gen.

Fonction d'appartenance (Fonc_App) : un SEF A dans un univers de discours X est caractérisée par sa Fonc_App μ(A)[x], qui associe à chaque item x (de X) une valeur réelle dans l'intervalle [0,1], mesure le degré auquel x appartient à l'ensemble A.

- ✓ Dans l'approche proposée, nous utilisons l'une des t-conormes les plus utilisées pour calculer le *Deg_App* de n-items : l'approche Max, définie par :

$$\mathrm{Max}_{i=1..n}\left(\mu(a_j)t[x_i]\right) = \perp_{j=1..m}\left(\mu(a_j)\right) = \mathrm{Max}_{j=1..m}\left(\mu(a_j)\right) \qquad \text{(IV.1)}$$

IV.3.2 Concepts de la logique possibiliste adaptés à notre approche

Introduite dans [71] comme une extension de la théorie des SEF, puis développée par [84], la théorie des possibilités est basée sur le calcul des mesures de possibilité et de nécessité.

Mesure de possibilité Π : pour un ensemble de référence X, Π assigne (affecte) à chaque sous-ensemble A ∈ X un nombre réel dans [0,1], évaluant à quel point un évènement A_i est possible, tel que : Π(∅)=0; Π(X)=1 ET $\forall A_i \in X,\ \Pi\big(\cup (A_i)\big) = Max\big(\Pi(A_i)\big)$
Cette fonction exprime que la possibilité de réaliser l'un des événements A_i, pris indifféremment, est la même que la réalisation de l'événement autant que possible entre eux.

Mesure de nécessité N : pour un ensemble de référence X, N indique à quelle mesure (degré) la réalisation d'un événement A_i est certaine. N est l'ampleur duale de Π. Cette mesure attribue à tous A un réel dans [0,1] et vérifie ensuite :
N(∅)=0; N(X)=1 ET $\forall A_i \in X, N(\cap (A_i)) = Min(\mathrm{N}(A_i))$

N peut être obtenu à partir de Π par : $\forall Y_i \in X, N(Y_i) = 1 - \Pi(Y_i^{\,c})$

- ✓ Dans notre approche, vu qu'on veut favoriser les expressions génétiques claires et fortes pour extraire les connaissances les plus pertinentes possibles, on a basé notre modélisation en logique possibiliste sur le calcul de la *mesure de possibilité*, qui correspond au t-conorme, c'est donc le choix du Max entre les Exp_Gen.

IV.3.3 Concepts de l'algorithme Apriori à adapter

Développé initialement pour l'analyse des tables de transactions dans le but de découvrir les d'articles achetés fréquemment ensemble, l'Apriori est l'algorithme de référence pour l'extraction des itemsets fréquents formant les RA binaires [54]. Cet algorithme n'est pas adapté au traitement de données nuancées. [67] définit la RA intégrant les concepts de la théorie des SEF. Une RA floue a la forme : « Si X est A », alors « Y est B »; formellement : (X,A) → (Y,B)

- ✓ Pour le calcul du support des itemsets, la méthode la plus adaptée à notre modélisation (basée sur les seuils) est : le somme_comptage_seuil, en additionnant uniquement les *Deg_App* (de chaque ligne t contenant cet itemset) dépassant un seuil w, soit :

$$Fuzzy\_supp(X,A) = \frac{\sum_{t\in T} \mathcal{T}(\mu(a_j)t[x_i])}{nb(T)} \qquad \text{(IV.2)}$$

Avec :

$\mathcal{T}$: norme triangulaire t-conorme utilisée pour calculer le *Deg_App* des itemsets flous
La cardinalité d'un itemset flou (via somme_comptage_seuil) : $Max_{i=1..n}(\mu(a_j)t[x_i]) > w$

- ✓ Ainsi, la formule du calcul de la confiance d'une RA floue sera donc :

$$Fuzzy\_conf((\mathrm{X},\mathrm{A}) \rightarrow (\mathrm{Y},\mathrm{B})) = \frac{Fuzzy\_supp(X \cup Y, A \cup B)}{Fuzzy\_supp(X,A)} \qquad \text{(IV.3)}$$

IV.4 Modélisation de l'approche proposée

Notre approche consiste à extraire des RA flous et possibilistes à partir de séquences d'images ISH d'expression génique. Il s'agit d'une critique et d'une amélioration de travaux antérieurs ayant tenté d'étudier ce même type de données [105] [100] [104]. Cependant, les auteurs ont tous manipulé ces données via une logique binaire inadaptée à leur réalité, ce qui entraîne une perte importante de connaissances biologiques cachées. Ainsi, après une série d'opérations de prétraitement, et dans le but de vectoriser les images ISH pour en extraire les caractéristiques locales, nous proposons une modélisation à la fois floue et possibiliste, représentant plus fidèlement la réalité des données étudiées, afin d'extraire des connaissances pertinentes et biologiquement plausibles.

IV.4.1 Le modèle en SEF proposé

Chacun des trois niveaux d'expression (fort (en rouge), modéré (en jaune) et non détecté (en gris)) sera ainsi représenté par un SEF et donc associé à une *Fonc_App* (Figure IV.1), ce qui affecte un *Deg_App* nuancé à chaque SEF. Cela implique une formulation du *Deg_App* selon la force d'Exp_Gen dans chaque case de 6x6 pixels de l'image ISH. Donc, il s'agit d'une extraction des caractéristiques de l'Exp_Gen que subit chaque case dans l'image.

La vectorisation des images étudiées, selon notre modélisation SEF, consiste à convertir ces images colorées en images en niveaux de gris (afin de définir l'Univers du Discours *Univ_Dis*). Les résultats des tests expérimentaux des niveaux de gris des pixels pour chacune des trois forces d'expression génétique ont donné les intervalles suivants des noyaux des trois *Fonc_App* selon leurs niveaux de gris : fort [0 - 76], non détecté [128 - 128] et modéré [225 - 255]. Par exemple, si le niveau de gris d'un pixel est égal à 100, il est considéré comme appartenant à un SEF-Fort avec un *Deg_App* de 0,5 et à un SEF-Non_Détecté avec un *Deg_App* de 0,5, sachant bien sûr que la somme de ces *Deg_App* doit être égale à 1.

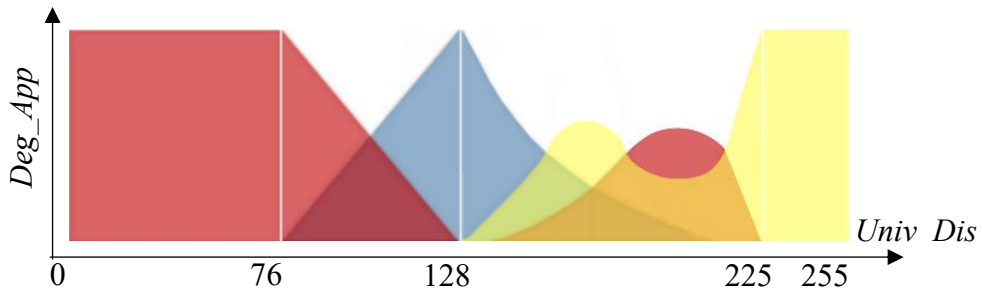


Figure IV.1 : Fonction d'appartenance définie pour la modélisation en SEF proposée (SEF_{Fort} en rouge, $SEF_{Modéré}$ en jaune, $SEF_{Non\_Détecté}$ en gris)

Ainsi, selon cette modélisation, toute case de 6x6 pixels dans les images ISH sera un attribut (une colonne dans la table des transactions), et chaque image sera une ligne. De plus, chaque cellule de cette table contiendra trois valeurs représentant les *Deg_App* d'une case à chacun des trois SEF, conformément à la *Fonc_App* de la figure IV.1 ci-dessus.

IV.4.2 Le modèle possibiliste proposé

Pour définir une fonction permettant de calculer le *Deg_App* à chaque SEF conformément au raisonnement possibiliste (qui n'exige pas que la somme des trois *Deg_App* soit égale à 1), nous avons étudié des estimateurs statistiques tels que le Maximum de Vraisemblance (Maximum Likelihood Estimation, MLE) et le Maximum A Posteriori (MAP). Le MAP est un estimateur d'un certain nombre de paramètres inconnus θ, tels que la densité de probabilité P relative à un échantillon X donné. Utilisé dans notre modélisation possibiliste, la formule de base de MAP est [118] :

$$\theta_{MAP}(X) = arg_{\theta}\, max\, \rho\,(X/\theta)\rho(\theta) \qquad \text{(IV.4)}$$

Ainsi, la classification par la méthode MAP recherche la classe la plus probable C_i (C_i=θ, un SEF pour nous) étant donné X (un vecteur de X_j), donc la classe qui maximise $P(X/C_i)$.

MAP est une méthode probabiliste basée sur la règle de Bayes :

$$\rho(C_i/X) = \rho(X/C_i) \times \rho(C_i)/\rho(X) \qquad \text{(IV.5)}$$

Avec :

$P(C_i/X)$ = probabilité d'appartenir à la classe C_i, X étant une case dans une image ISH

$P(X/C_i)$ = probabilité d'appartenir de la case X (de chacune de ses 4 zones X_j) à la classe C_i (la méthode MAP choisira le maximum)

$P(C_i)$ = probabilité d'appartenir à la classe C_i (indépendant)

P(X) = probabilité d'être la case X

Pour chaque zone X_j, $P(X_j)$=1/4 (une constante). Par conséquent, $P(C_i)/P(X)$ est constant pour tout X_j. Ainsi, $P(C_i/X)$ ne dépend que de $P(X/C_i)$, alors :

$$\rho(C_i/X)=MAP(\rho(X_j/C_i)) \qquad \text{(IV.6)}$$

La vectorisation possibiliste proposée est la suivante : on divise chaque case X de 6x6 pixels en 4 zones X_j de 3x3 pixels. Ensuite, selon la méthode MAP, on calcule pour chaque zone X_j la probabilité d'appartenir à chacun des trois SEF ; le maximum entre ces 4 probabilités sera le *Deg_App* de cette case X dans l'image (cellule dans la table) à un SEF respectivement.

IV.4.3 Pseudo-algorithme de l'approche proposée

Par conséquent, lorsque le *Deg_App* d'une zone (et donc d'une case) est égal à 1, cette case est dite homogène, car elle contient une zone ayant subi une Exp_Gen claire dans l'un des trois niveaux d'expression. L'organigramme suivant illustre les opérations de prétraitement effectuées et la modélisation des incertitudes proposée pour extraire les deux tables de transactions et en déduire les RA les plus pertinentes (RA floues ou RA possibilistes).

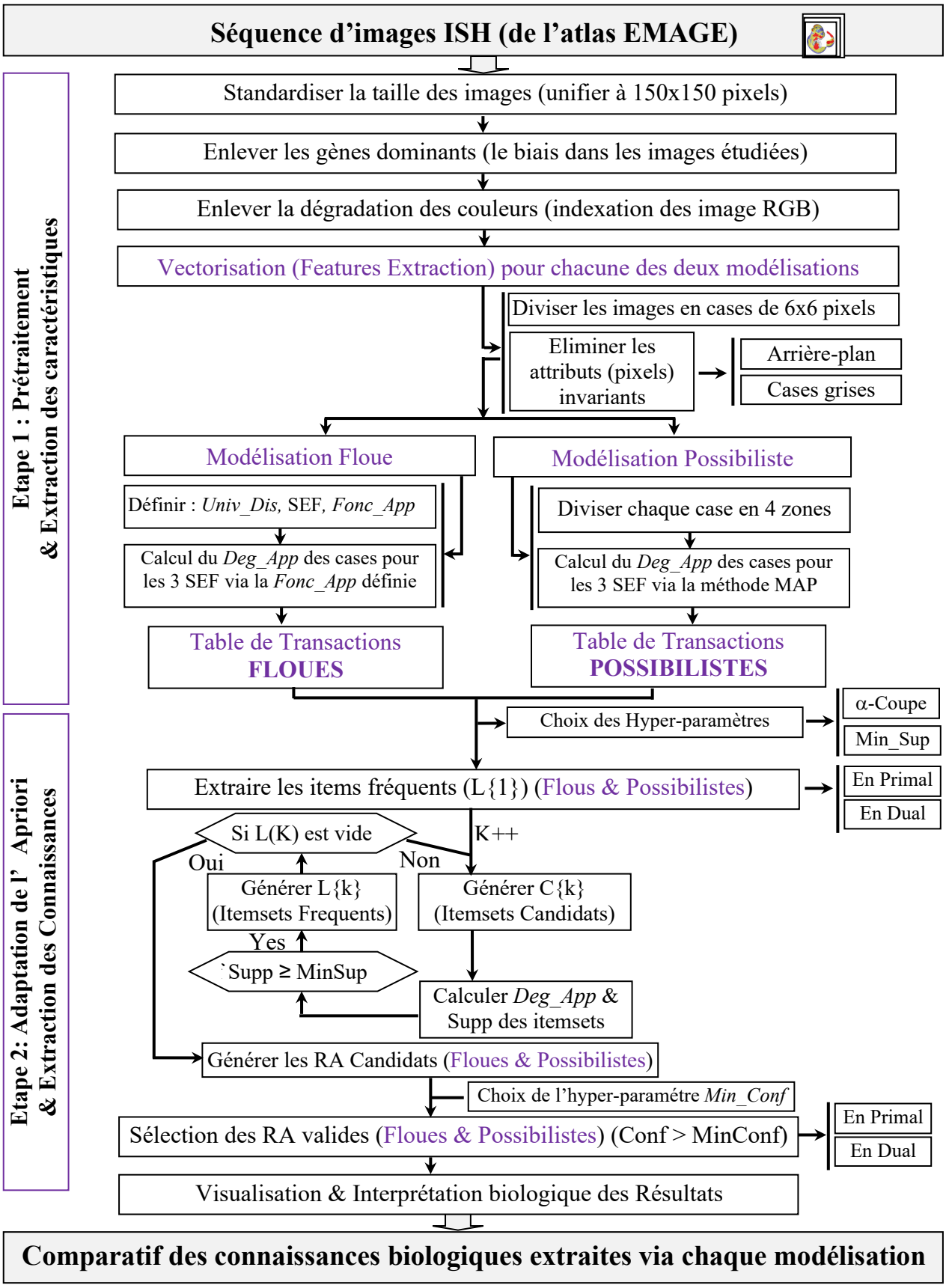


Figure IV.2 : Pseudo-Algorithme des deux approches proposées

IV.4.4 Contribution

Notre approche vise à proposer une modélisation adaptée à la réalité des données étudiées, garantissant l'extraction de connaissances cachées ayant une signification biologique claire. Contrairement aux travaux précédents [105] [100] [104], qui traitaient les images d'Exp_Gen suivant une logique binaire, en réduisant les formes d'expression aux seules formes *détectées* ou *non détectées*, et en négligeant toutes les autres formes nuancées d'expression possibles, cela éliminerait des détails très intéressants et entraînerait une perte importante de connaissances biologiques cachées dans ces images. Ainsi, nous avons étudié les trois formes d'Exp_Gen les plus significatives (forte, modérée et non détectée) afin de conserver les connaissances les plus pertinentes. Les autres formes d'expression (faible et possible) ne font qu'accroître la complexité algorithmique de l'algorithme Apriori (initialement élevée) sans apporter des connaissances utiles.

Par ailleurs, bien que le dataset EMAGE fournisse des images de haute qualité, nous ne pouvons ignorer le risque d'imprécision lors de l'annotation spatiale des Exp_Gen sur ces images, en supposant avoir cerner exactement les limites des zones colorées (et donc les niveaux d'Exp_Gen). Aussi, les gènes dominants peuvent engendrer des résultats biaisés ; nous les avons retirés de l'étude pour éviter tout biais dans les connaissances extraites.

Les opérations de prétraitement réalisées ont permis de réduire nettement la dimensionnalité des données étudiées, tout en conservant une variabilité maximale de leur contenu. Pour l'extraction de caractéristiques, la modélisation proposée (floue et possibiliste) a permis de traduire fidèlement la réalité de ces données via leur vectorisation vers deux tables de transactions, ceci par :

- Compression de chaque image de 150x150 pixels en une image de 25x25 cases nuancés de 6x6 pixels (passant de 150^2 à 25^2 variables, soit une réduction de dimension de 97,3%).
- Suppression des attributs invariants non porteurs de connaissances, représentant : les cases blanches (l'arrière-plan), les cases toujours grises (n'ayant subi aucune Exp_Gen dans les 1694 images étudiées) ; et enfin la suppression des attributs (colonnes) ayant subi une Exp_Gen faible (ceux ayant toujours un Deg_App inférieur au α-coupe de la Fonc_App).

Aussi, pour modéliser plus fidèlement la réalité (loin d'être binaire et précise) des données biologiques étudiées :

- Le modèle en SEF proposé fournit une affectation nuancée de chaque case de 6x6 pixels de l'image ISH à chacune des trois formes d'expression étudiées via une *Fonc_App* pour chacun des 3 SEF.
- Nous avons étendu notre approche en integrant la logique possibiliste pour résoudre le problème de l'effet de flou souvent créé artificiellement suite à l'affectation nuancée de chaque case via la modélisation en SEF. Cet effet efface les limites entre les zones

d'Exp_Gen, mettant à égalité toute forme de distribution (concaténée ou dispersée) des pixels colorés dans une case. Ainsi, grâce à la méthode MAP, les limites des zones d'expression seront préservées et les zones ayant subi une Exp_Gen claire (pixels concaténés de même couleur) seront favorisées.

- De plus, le calcul du Deg_App par le MAP permet de resoudre le problème des pixels bruités (ayant une couleur non homogène avec les autres pixels adjacents) en privilégiant les pixels adjacents les plus homogènes.

Globalement, nos deux modélisations permettent de découvrir trois formes différentes de RA (et donc de connaissances biologiques), selon deux points de vue (Primal et Dual) :

- Extraire les RA entre zones d'Exp_Gen (relations spatiales) : lorsqu'un ensemble de pixels adjacents (donc un organe) subit une Exp_Gen, quels autres organes sont en relation biologique avec ce dernier, et devraient également subir une Exp_Gen ; si c'est le cas, cela serait selon quelle force d'expression ?
- Extraire les RA entre les gènes : lorsqu'un gène est exprimé, quels autres gènes doivent également s'exprimer (formuler les relations de co-expression génétique) ?
- À partir des RA entre gènes, et puisque nous avons utilisé une nomination combinant le nom de chaque gène avec le Theiler Stage (la phase de dévelopement embryonique) dans lequel ce gène a été exprimé, une nouvelle forme de connaissance temporelle sera extraite, nous aidant à comprendre les relations entre les Exp_Gen durant les phases de développement des embryons des espèces vivantes. Ce type de RA extraites est très utile puisque ces relations détectées aident à formuler clairement l'aspect spatio-temporel dans les données étudiées.

Malgré l'explosion combinatoire des sous-ensembles (itemsets) possibles de gènes, les résultats obtenus montrent qu'avec un MCT de 90%, le taux de relations génétiques (dans les RA extraites) reconnues par la plateforme String-DB.org [119] est de 37,74% pour les RA floues et de 42,56% pour les RA possibilistes. Enfin, 62,94% des RA floues présentent des relations temporelles entre les mêmes phases (même TS), et 37,06% entre différentes TS. Pour les RA possibilistes, 77,81% sont entre les mêmes TS, et 22,19% entre différentes TS.

Partie 2 : Implémentation de l'approche & Interprétation biologique des résultats

IV.5 Spécificités des données d'application

IV.5.1 Pourquoi limiter l'étude à une sous-partie du DataSet Emage

Le dataset étudié, EMAGE, est une collection d'images ISH de haute qualité, qui vise à capturer : la localisation spatiale (zones) des niveaux d'Exp_Gen dans la morphologie de l'embryon (tissus, organes ou corps entier) ; et la phase temporelle d'expression des gènes, appelée phase de Theiler (Theiler Stage, abrévié TS). Ainsi, EMAGE intègre divers modèles spatiotemporels cachés d'expression génétique [10].

Pour des raisons techniques et économiques, la souris est devenue de loin l'animal le plus utilisé par les chercheurs pour étudier l'impact de nouvelles thérapies médicamenteuses et génétiques en les appliquant sur plusieurs générations de cette espèce modèle, compte tenu de la rapidité de croissance de son embryon (de l'ordre de 19 à 23 jours, les Theiler Stage). De plus, le génome de la souris est également entièrement séquencé et est très proche de celui de l'homme [10].

Enfin, et pour des contraintes techniques liées à l'approche proposée, nous avons cerné l'étude à 1694 images (en format PNG) d'expression génétique couvrant les TS16, TS17 et TS18. Puisque durant ces trois phases de développement de l'embryon, le corps ne subit pas de grandes déformations, cela permet une comparaison spatiale entre les zones d'Exp_Gen.

IV.5.2 Eliminer le biais (les gènes dominants) dans le dataset étudié

Parmi les images d'Exp_Gen fournies par le dataset Emage figure un type de gènes qui présentent une forme d'expression (rouge ou jaune) touchant une partie très vaste (la majorité des pixels) de l'image de l'embryon. Ce sont les images représentant les Exp_Gen des *gènes dominants*, qui peuvent engendrer des résultats biaisés lors de l'extraction des connaissances, puisqu'ils seront présents (imposés) dans un grand nombre de RA spatiales extraites. Ces RA seront donc triviales (évidentes) et non pertinentes et n'ont aucun intérêt biologique.

On peut distinguer deux types d'images de gènes dominants :

- Images presque entièrement en rouge (Whole Red) : ayant subi une forte Exp_Gen dans de grandes zones.
- Images presque entièrement en jaune (Whole Yellow) : ayant subi une Exp_Gen moyenne dans de grandes zones.

Ainsi, les gènes dominants peuvent engendrer des résultats biaisés ; nous avons retiré leurs images du dataset étudié pour éviter tout biais dans les connaissances extraites.

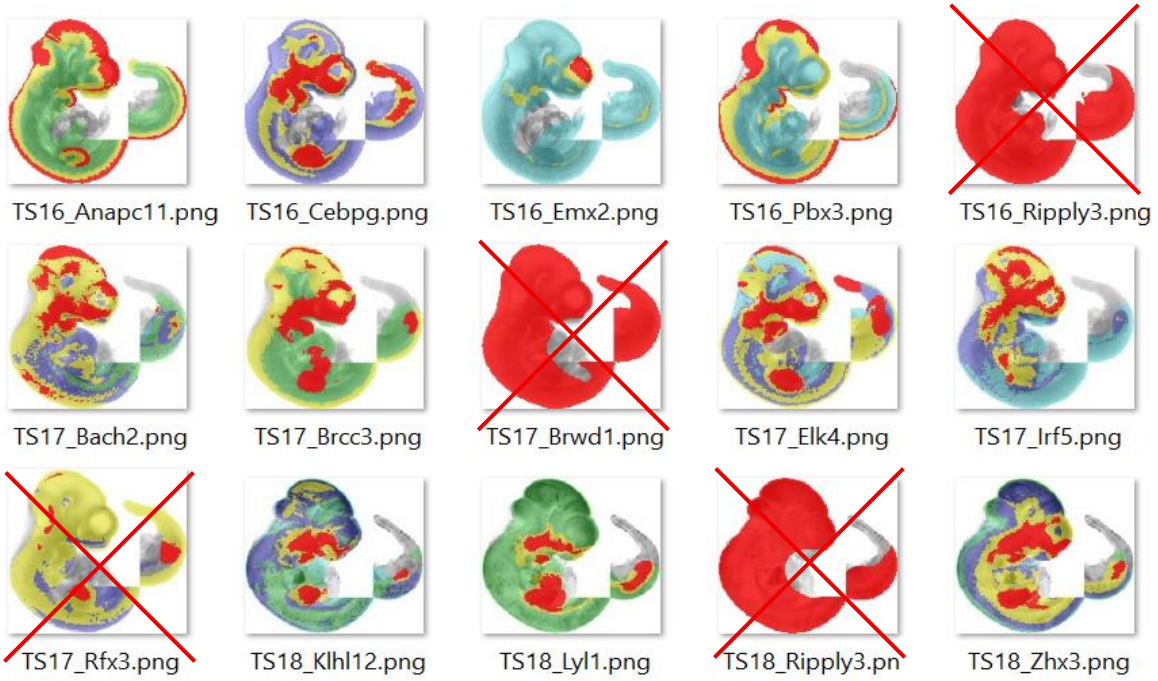


Figure IV.3 : Exemples des types de gènes dominants à éliminer (whole red, whole yellow)

IV.5.3 Outil de validation biologique des résultats : la plateforme String-DB

STRING-DB (*Search Tool for the Retrieval of INteracting Genes/proteins - Data Base*) est un dataset des interactions protéines-protéines (ou gènes-gènes) connues et prédites par la communauté des chercheurs dans le domaine de la biologie. Les interactions comprennent des associations directes (physiques) et indirectes (fonctionnelles) qui découlent de la prédiction computationnelle, du transfert de connaissances entre organismes et des interactions agrégées à partir d'autres bases de données (primaires) [site : String-DB.org] [119]. Les interactions dans STRING-DB sont dérivées de cinq sources principales :

Figure IV.4 : Les principales sources biologiques des données de String-db.org [119]

Les jeux de données agrégés au sein de la plateforme STRING-DB couvrent actuellement 24.584.628 protéines, issues de 14.094 organismes vivants, offrant ainsi une couverture exceptionnellement large des interactions protéiques à travers l'ensemble des espèces modèles.

IV.6 Vectorisation des images ISH via les deux modélisations

IV.6.1 Etapes du processus de vectorisation

La vectorisation consiste à convertir (suivant chacune des deux modélisations : floue, possibiliste) l'ensemble des 1694 images d'Exp_Gen étudiées en deux fichiers CSV appelés *tables de transitions*. Ainsi, toutes les images acquissent par les capteurs dans leur forme originale (dataset Emage) durant les TS16-TS17-TS18 vont subir les mêmes étapes de prétraitement détaillées dans la Table Tab. IV.1. Ces étapes sont comme suite :

- *Redimensionner* les images PNG collectionnées pour unifier leur taille, on obtiendra des images de taille standard de 150x150 pixels.
- *Indexer* chaque image PNG (initialement acquise en mode RVB) en 4 couleurs : Rouge (expression forte), Jaune (moyenne), Gris (expression inexistante), Blanc (l'arrière-plan).
- *Unifier* (rendre à gris) la couleur des *pixels non porteurs de connaissances utiles* : pixels à couleurs grise et blanche (donc éliminer l'arrière-plan).
- *Diviser* chaque image résultante *en cases* de 6x6 pixels chacune, donc chaque image contiendra 25x25 cases.
- *Extraction des caractéristiques* (Features Extraction) pour obtenir une table de transactions selon chaque type de modélisation (floue et possibiliste). Chaque case dans l'image (qui sera aussi une case dans chacune de ces deux tables) sera formée de trois cellules, chacune contiendra le *Deg_App* de cette case à chacun des trois SEF suivant chaque logique. Les valeurs des cellules ans chaque case seront calculées comme suite :
    - ***Via la modélisation en logique floue :*** on calcule les *Deg_App* de chaque case à chaque SEF suivant la Fonc_App définie dans la figure IV.1. Ce *Deg_App* représente aussi le pourcentage des pixels de chacune des trois couleurs (force d'Exp_Gen) parmi les 6x6 pixels d'une case.
    - ***Via la modélisation en logique possibiliste :*** basée sur la méthode MAP (Maximum A Postériori). On divise chaque case (de l'image) en 4 zones, chaque zone sera constituée de 3x3 pixels. Chacune des cases (de la table) est formée de 3 cellules, et chaque cellule représente la *possibilité* pour que cette case soit de couleur rouge, jaune ou grise. On calcule la possibilité d'appartenance de chaque zone à chacun des 3 SEF, ensuite via la méthode MAP on choisit (pour chaque SEF) la possibilité maximale entre celles de 4 zones (correspondante à la norme triangulaire t-conorme). La méthode MAP se base sur la règle de Bayes (voir l'équation (III.1), donc : $P(C_i/Z) = P(Z/C_i) \times P(C_i)/P(Z)$ avec :
      P : Probabilité C_i : La classe i (Couleur) Z : La zone

Ainsi, chaque ligne des deux tables issues de chacune des deux modélisations représente un gène (une image), c'est la vectorisation de l'image d'expression génétique de ce gène suivant

chacune des deux logiques de modélisation proposées. Aussi, chaque colonne représente une case (i,j) ans l'image. Finalement, chacune de ces deux tables sera l'entrée (input) de l'algorithme Apriori (adapté à ces deux logiques de modélisation) pour extraire des RA entre cases (RA spatiale) et entre gènes, cela via deux points de vue : primale, duale (voir Tab. IV.1).

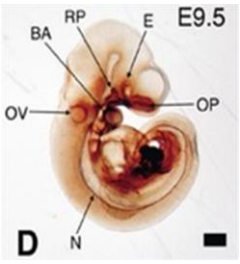
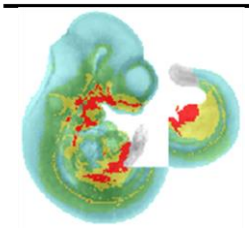


Image annotations :
BA : Branchial Arch
E : Eye
N : Notocord
OP : Olfactory Placode
OV : Otic Vesicle
RP : Rathke's Pouch
Size bar= 200um

La base EMAGE des images d'expression génétique est fournie dans le lien www.emouseatlas.org/emage. Notre étude a été limitée à 1694 images des Theiler Stages 16-17-18 puisque la forme de l'embryon est généralement identique durant ces trois phases, ce qui permettra une bonne comparaison spatiale entre les zones d'Exp_Gen.

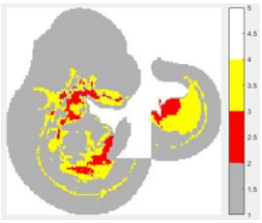
L'image à gauche est un exemple d'une image originale (durant l'acquisition) d'un gène nommé *Acvr1* du Theiler Stage TS16 (age = 9.5 dpc (days post coitum)). (Reference: EMAGE:3259 dans le database online)



Annotation color key:
Forte
Modérée
Faible
Possible
Non Détectée

Annotation Spatiale : après l'acquisition de l'image, une opération de *mappage de couleurs* sera effectuée, et les zones d'expression génétique seront colorées en accordance avec la force de l'expression génétique.

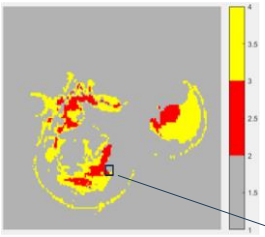
Redimensionnement : lors de l'acquisition, les dimensions des images étudiées sont différentes, donc nous avons opté à une standardisation des dimensions de ces images à 150x150 pixels.



Annotation color key:
Fort
Modérée
Non Détectée
Arrière-Plan

Indexation : c'est l'étape de transition d'une image colorée en mode RVB vers une image indexée en 4 couleurs : rouge (Exp_Gen fort), jaune (modérée), gris (Exp_Gen : faible, possible ou non-détectée), et finalement le blanc (l'arrière-plan).

Cette étape aidera à éliminer les dégradations dans la même couleur, qui rendra possible la classification automatique des pixels dans l'un des trois niveaux d'Exp_Gen étudiés.

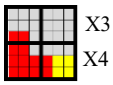


Ex : case N° 476

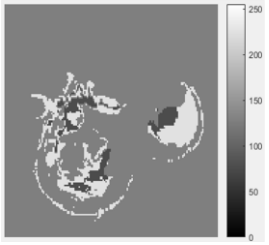
Unifier la couleur des pixels invariants (pixels non porteurs de connaissances) à la couleur grise. Les pixels invariants sont :
- Les pixels blanc (d'arrière-plan)
- Les pixels gris (n'ayant subi aucune Exp_Gen)

Extraction des caractéristiques : après avoir diviser les images en cases de 6x6 pixels, on passe aux *features extraction de* chaque de case suivant l'exemple au-dessous et via chacune des deux modélisations proposées :

Modélisation Floue



Modélisation Possibiliste



SEF Fort	SEF Modéré	SEF Non-détecté
P(Rouge) = 12/36	P(Jaune) = 4/36	P(Gris) = 20/36
Deg_App du SEF Fort	*Deg_App* du SEF Modéré	*Deg_Ap* du SEF Non détecté
R=0.33	**J=0.11**	**G=0.56**

SEF Fort	SEF Modéré	SEF Non-détecté
P(X1/Fort)=2/9 P(X2/Fort)=8/9 P(X3/Fort)=0/9 P(X4/Fort)=2/9	P(X1/Modéré)=0/9 P(X2/Modéré)=0/9 P(X3/Modéré)=0/9 P(X4/Modéré)=4/9	P(X1/Non-Det)=7/9 P(X2/Non-Det)=1/9 P(X3/Non-Det)=9/9 P(X4/Non-Det)=3/9
MAP du SEF Fort	MAP du SEF Modéré	MAP du SEF Non détecté
R=0.89	**J=0.44**	**G=1**

Calculer le *Deg_App* à chaque SEF via le *Modèle Flou* proposé. La cellule générée est :
[476, R, 0.33] ; [476, J ,0.11] ; [476, G, 0.56]

Calculer le *Deg_App* à chaque SEF via le *Modèle Possibiliste* proposé. La cellule générée est :
[476, R, 0.89] ; [476, J, 0.44] ; [476, G, 1.00]

Chaque case dans l'image = Cellule dans la table
Chaque cellule = trois éléments
Un élément = [N°case, SEF, *Deg_App]*

Table IV.1 : Exemple du processus de vectorisation : sequence d'operations de pretraitement pour l'extraction des tables de transaction Floues & Possibilistes à partir des images ISH étudiées

IV.6.2 Exemples de tables résultantes de la vectorisation des images ISH

Ainsi, chaque image du dataset Emage, représentant l'expression d'un gène, qui deviendra une ligne dans les tables de transaction résultantes, a subi le même processus de prétraitement décrit dans la table Tab. IV.1. Les deux tables issues du processus de prétraitement proposé ont la structure suivante :

- Chaque ligne (transaction i) représente la vectorisation d'une image d'Exp_Gen, nous aurons donc 1694 lignes.
- Chaque colonne (attribut j) représente une case de l'image (6x6 pixels), donc 625 colonnes.
- Chaque case (i, j) dans la table, est composé de trois cellules, chacune représente le *Deg_App* de cette case à l'un des trois SEF (voir l'exemple à la fin de la table IV.1).

Lors de l'implémentation de notre approche, et pour réduire encore plus la complexité algorithmique, on a supprimé les attributs représentant le SEF Non-Détecté puisqu'il est non porteur de connaissances. Puis, on a éclaté la cellule en deux attributs, chacun représente l'Exp_Gen en cette case suivant les deux SEF Rouge et Jaune. Ainsi, la case sélectionnée dans chacune des deux tables au-dessous (Tab. IV.2 et Tab. IV.3) représente une case dans d'une image d'expression génétique. Elle est formée de trois cellules, contenant les *Deg_App* à chacun des deux SEF significatifs, suivant la logique de modélisation utilisée.

[1 0.8889 36]	[2 0.1111 36]	[1 0.5833 37]	[2 0.1389 37]
[1 0.4722 36]	[2 0.1667 36]	[1 0.4167 37]	[2 0.0556 37]
[1 0.9167 36]	[2 0 36]	[1 0.3611 37]	[2 0 37]
[1 0 36]	[2 0.6111 36]	[1 0 37]	[2 0.3889 37]
[1 0.7222 36]	[2 0.2778 36]	[1 0.7500 37]	[2 0 37]
[1 0.3333 36]	[2 0.1667 36]	[1 0.3333 37]	[2 0.1389 37]
[1 0.9722 36]	[2 0 36]	[1 0.5556 37]	[2 0 37]

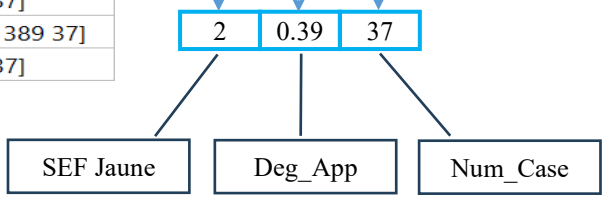


Table IV.2 : Extrait de la table de transactions issue de la modélisation en logique Floue

[1 0 33]	[2 0.3333 33]	[1 0 34]	[2 0.6667 34]
[1 0.7778 33]	[2 1 33]	[1 1 34]	[2 0.8889 34]
[1 0 33]	[2 0.2222 33]	[1 0 34]	[2 1 34]
[1 0 33]	[2 0.3333 33]	[1 0 34]	[2 0.6667 34]
[1 0 33]	[2 0.5556 33]	[1 0 34]	[2 0.7778 34]
[1 0.5556 33]	[2 0.8889 33]	[1 0.3333 34]	[2 0.8889 34]
[1 1 33]	[2 0.7778 33]	[1 1 34]	[2 0.6667 34]

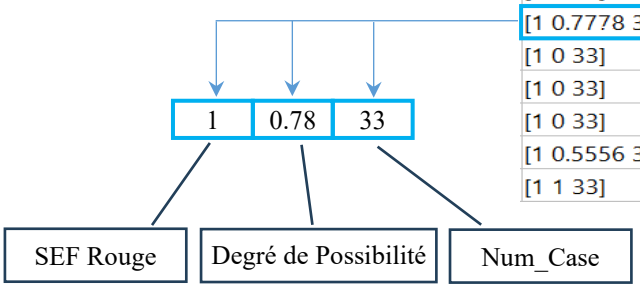


Table IV.3 : Extrait de la table de transactions issue de la modélisation en logique Possibiliste

On remarque que la somme des possibilités d'appartenance dans la case sélectionnée de la table Tab. IV.3 (qui est : 0 +0.78 + 1 = 1.78) est supérieure à 1, c'est l'une des caractéristiques principales de la logique possibiliste, qui nous libère de la condition de base dans les logiques antérieures (binaire, probabiliste et floue), qui impose que : Σ Deg_App = 1.

IV.6.3 Réglage des paramètres et Génération des tables de transaction

Nous avons effectué une diversité de tests empiriques dans le but de choisir des valeurs adéquates pour les hyperparamètres suivants :

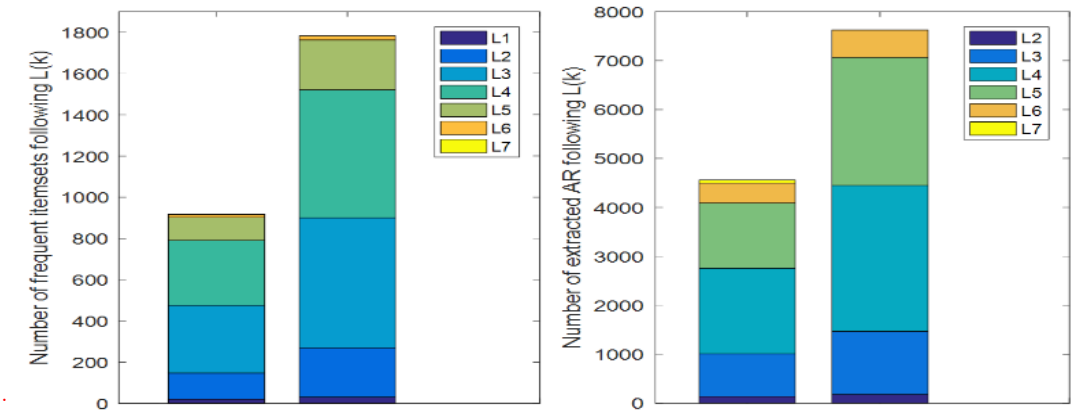
- α-coupe : pour conserver les attributs (items) les plus pertinents, nous avons choisi (pour les *Fonc_App* des trois SEF) une α-coupe ω = 0.3 comme seuil minimal de *Deg_App*.
- Min_Sup : pour sélectionner les itemsets fréquents, on a fixé le seuil Min_Sup à 0.3 pour les itemsets flous, et Min_Sup = 0.43 pour les itemsets possibilistes. Cela nous a aidé à éviter une explosion combinatoire des itemsets possibles, dont la majorité sont non fréquents.
- Min_Conf : pour filtrer les RA les plus pertinentes, on a levé le seuil Min_Conf à 0.9 .

IV.7 Analyse et interprétation biologique des résultats

IV.7.1 Analyse et discussion sur les RA spatiales extraites en *Primal*

Globalement, les RA extraites représentent : les relations spatiales (RA entre les cases), les relations temporelles entre les phases de développement de l'embryon (Theiler Stages (TS)) et les relations entre les gènes qui co-expriment.

Les histogrammes ci-dessous visualisent les itemsets fréquents générés, suivant leur L(K), ainsi que les RA extraites via les deux approches proposées.



Floue MST = 0.3 | Possibiliste MST = 0.37 | Floue MST = 0.3 | Possibiliste MST = 0.43

Figure IV.5 : Histogramme des itemsets Flous et Possibles fréquents et RA extraites en Primal

À partir de ces résultats, nous remarquons que la modélisation floue génère un nombre limité d'itemsets fréquents, et donc peu de RA floues seront générées. Mais, la modélisation possible a généré plus d'itemsets fréquents, et donc plus de RA seront extraites. Ceci peut être expliqué par le principe du MAP, qui choisit pour chaque case le *Deg_App* maximal à un SEF entre les *Deg_App* de ses quatre zones, ce qui génère plus d'itemsets fréquents, et donc de RA.

De plus, certaines RA floues ou RA possibles ont des lifts > 1, ça s'explique également par les spécificités du MAP et du t-conorme utilisés dans nos deux modélisations, qui choisissent toujours le maximum de *Deg_App* dans chaque case.

	Antecedent	Consequent	Supp	Conf	Lift	Distance Euclidienne
RA Floue	F_233 ^ F_510	F_263	0.30	0.96	2.16	11.18
	F_484	F_263 ^ F_485 ^ F_510	0.30	0.93	2.73	10.44
RA Possibiliste	F_237 ^ F_510	F_262 ^ F_263	0. 37	0.98	1.85	11.18
	F_233 ^ F_258 ^ F_485	F_234	0.38	0.99	2.10	10.20

Table IV.4 : Exemples de RA Floues & Possibilistes en Primal

Pour favoriser les RA qui sont les plus porteuses de nouveautés et de connaissances biologiques pertinentes, nous filtrons les RA triviales, résultantes des :

- Cases adjacentes : Étant donné que la zone d'Exp_Gen peut toucher plusieurs cases adjacentes dans l'embryon, cela provoquera donc l'extraction des RA entre ces cases adjacentes qui forment en réalité le même organe, et la relation entre ces cases sera ainsi évidente et inutile à extraire. On a donc choisi les RA ayant la plus grande distance euclidienne entre les cases (items) qui forment l'itemset de cette RA.
- RA chevauchées : ce sont les RA qui ont des sous-itemsets communs à la fois dans leurs antécédents et les conséquents. Ce n'est qu'une forme inutile de RA redondantes.

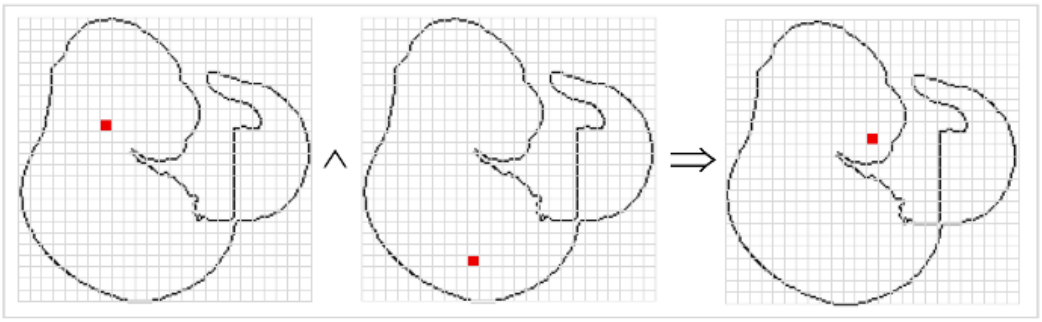


Figure IV.6 : Exemple de RA "Spatiale" extraite (entre les cases (organes))
(RA-Floue en Primale : Fort_233 ^ Fort_510 ⇒ Fort_263)

IV.7.2 Intérêt de la dualité dans l'approche proposée

Suivant le principe de la "dualité", nous pouvons générer deux nouvelles tables, qui sont les transposées de nos deux tables de transaction initiales, où une colonne (attribut) représentera un cette fois un gène (image d'Exp_Gen) et la ligne représentera les cases des images étudiées.

Par conséquence, cette dualité permet l'extraction de RA cruciales représentant les relations entre les gènes. De plus, étant donné la nomination combinée [TS-Gène] des images d'Exp_Gen qu'on a utilisé dans notre approche, ces RA sous forme de Genes → Genes clarifieront le séquençage temporel des Exp_Gen pendant les phases de développement embryonnaires (voir les tables Tab. VI.5 et Tab. VI.6).

	Antecedent	Consequent	Supp	Conf	Lift
RA Floues	TS16_Zfp410_Fort	TS17_Zfp623_Fort	0.36	1.10	3.33
	TS17_Skp1a_Modéré	TS16_Zfp410_Fort	0.30	0.99	2.74
		TS17_Zfp623_Fort			
RA Possibilistes	TS17_Zfp623_Fort	TS16_Zfp410_Fort	0.48	0.97	1.92
	TS16_Nfatc2_Modéré	TS17_Zfp623_Fort	0.43	0.90	1.82
	TS16_Zfp410_Modéré				

Table IV.5 : Exemple de RA Floues & Possibilistes en "Dual"

La table suivante (Tab. VI.6) représente un exemple de RA entre gènes, durant les deux phases de développement TS16 et TS17, extraites via le principe de la dualité.

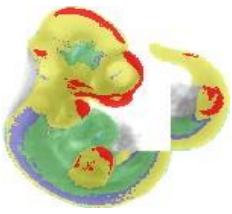
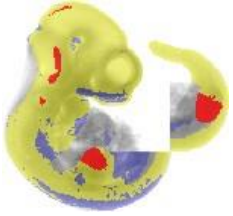
Antecedents		Consequents		
TS17_Skp1a Modéré	⇒	TS16_Zfp410 Fort	TS17_Rfx3 Modéré	TS17_Zfp623 Fort
				

Table IV.6 : Exemple graphique de RA-Floues en "Dual"
(MST=0.28, MCT=0.93, Lift=2.89)

IV.7.3 Interprétation biologique des résultats obtenus en *Dual*

IV.7.3.1 Vision globale des résultats de la modélisation floue proposée

La figure ci-dessous représente le graphe de dépendances entre les gènes suivant la modélisation en logique floue proposée. Ce graphe est généré via la plateforme String-DB.org pour analyser les résultats de notre modélisation floue. En analysant ce graphe global de co-expression génétique résultant des itemsets flous fréquents détectés via l'algorithme Apriori, on remarque que :

- La plateforme String-DB.org a reconnu la liaison entre une bonne partie des gènes formant ce graphe, 20 gènes ayant une liaison reconnue parmi un total de 53 gènes, ce qui fait un taux de 38%. De plus, la majorité de ces liaisons sont fortes, puisque la force de liaison est représentée par un nombre élevé d'arcs entre les nœuds (les gènes) du graphe. Cela prouve l'efficacité de l'approche proposée ainsi que la pertinence et la crédibilité de ses résultats.
- Ainsi, une partie importante des gènes qu'on a détecté leurs co-expressions via notre approche de Machine Learning sont en cohérence avec les co-expressions déjà connues par String-DB. De plus, les 62% de gènes ayant une co-expression non signalée (non détectée) par String-DB représentent des liaisons qui ne sont pas *encore reconnues* par cette plateforme (donc des co-expressions inconnues par les chercheurs), et ces co-expressions sont potentiellement celles qui portent les nouveautés biologiques qui intéressent les biologistes, ce qui est le plus important dans notre travail.

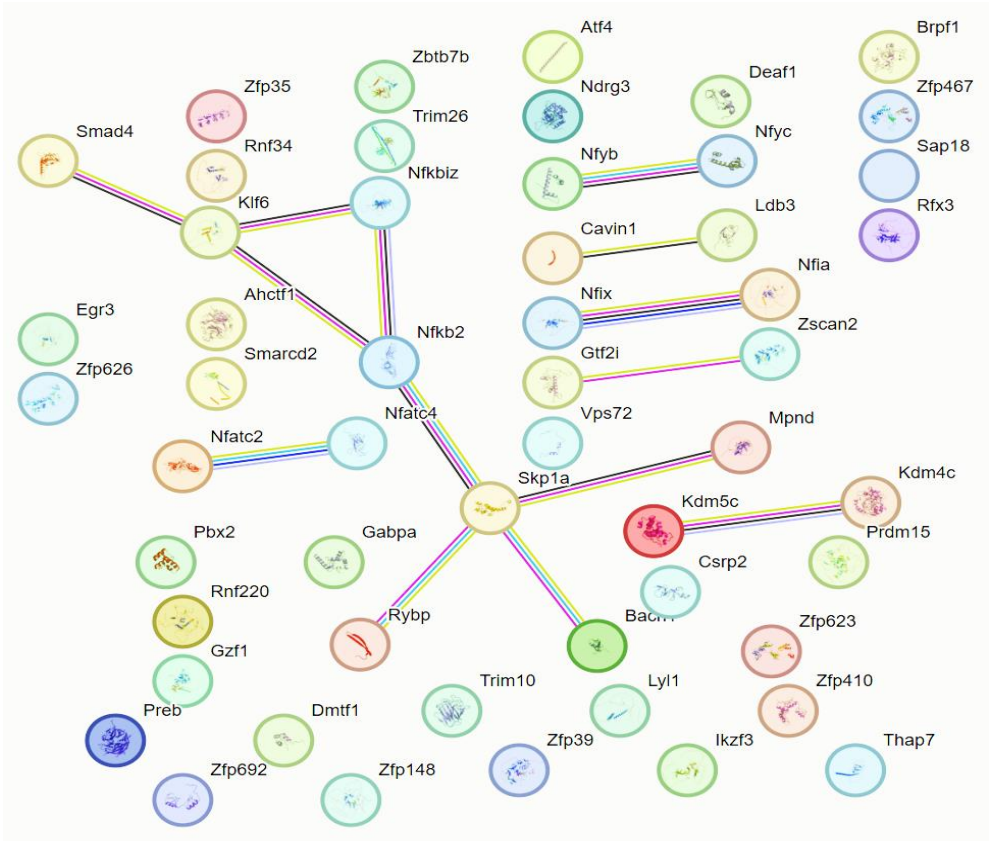


Figure IV.7 : Graphe de dépendances entre les gènes formant les itemsets flous fréquents en dual (généré par la plateforme String-DB.org)

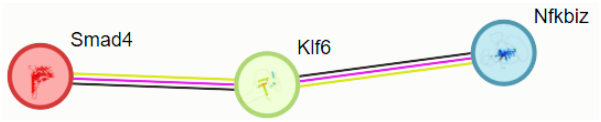
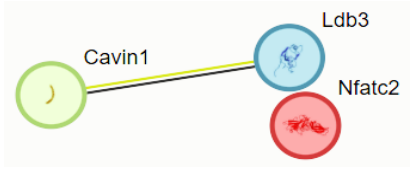
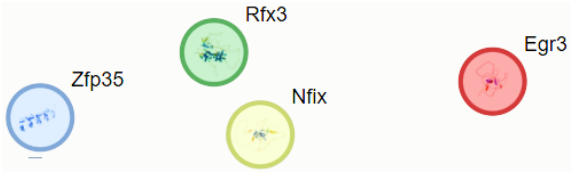
Exemple d'itemset extraits et reconnus *Totalement* par la plateforme String-DB

Exemple d'itemset extraits et reconnus *Partiellement* par la plateforme String-DB

Exemple d'itemset extraits et *Non Reconnus* par la plateforme String-DB.org


Table IV.7 : Exemples d'itemsets Flous et leur taux de reconnaissance par String-db.org

IV.7.3.2 Vision globale des résultats de la modélisation possibiliste proposée

La figure ci-dessous représente le graphe de dépendances entre les gènes suivant la modélisation possibiliste proposée. Ce graphe est également généré via la plateforme String-DB.org. Pour éviter la complexité algorithmique de l'Apriori, nous avons été contraints de lever le Min_Sup de la modélisation possibiliste à 43%, puisque cette dernière a découvert des centaines d'itemsets fréquents, et des milliers de RA à confiance élevée. En analysant ce graphe on remarque que :

- La plateforme String-DB.org a reconnu la liaison entre 43% des gènes formant les itemsets fréquents. Si String-DB n'a pas signalé une co-expression pour 57% des gènes, cela ne veut pas dire que nos résultats sont erronés (ces résultats sont supportés par des métriques mathématiques solides (support, confiance et lift assez élevés), cela signifie seulement que ces liaisons ne sont pas *encore reconnues* par String-DB.
- Ces itemsets fréquents peuvent être porteurs de nouveautés biologiques pas encore connues. Cela pourrait fournir un domaine de recherche « assez cerné » pour les biologistes, afin de confirmer ces relations encore dissimulées entre les sous-ensembles de gènes qui peuvent co-exprimer durant les phases de développement des vivants.

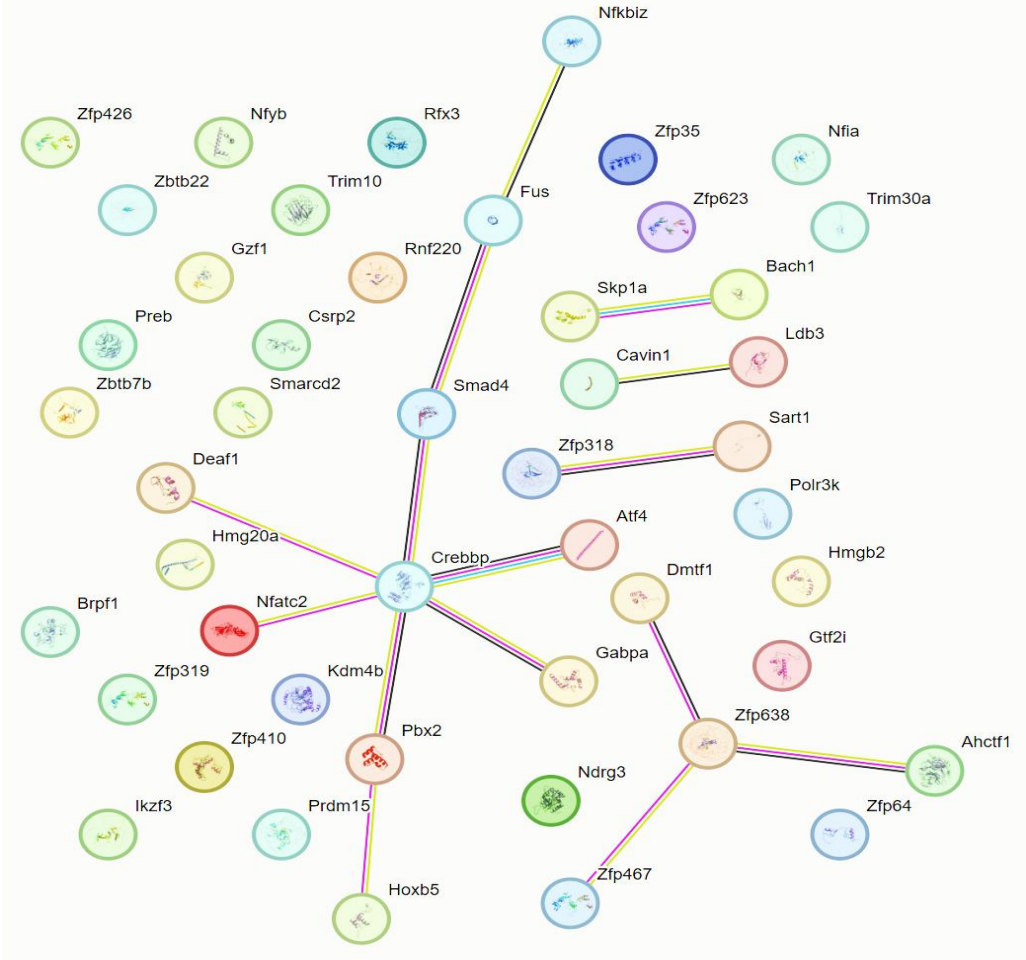


Figure IV.8 : Graphe de dépendances entre les gènes formant les itemsets possibilistes fréquents en dual (généré par la plateforme String-DB.org)

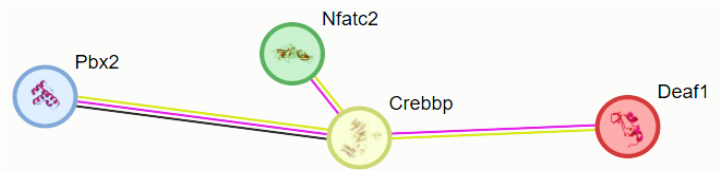
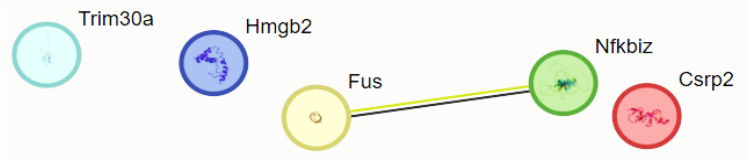
Exemple d'itemset extraits et reconnus *Totalement* par la plateforme String-DB

Exemple d'itemset extraits et reconnus *Partiellement* par la plateforme String-DB

Exemple d'itemset extraits et *Non Reconnus* par la plateforme String-DB.org


Table IV.8 : Exemples d'itemsets Possibilistes et leur taux de reconnaissance par String-db.org

IV.8 Conclusion

Après avoir présenté le cadre théorique et les détails de modélisation de l'incertain introduits dans l'approche proposée, nous avons détaillé les adaptations de l'algorithme Apriori nécessaires pour extraire des RA suivant la logique floue et possibiliste. Ensuite, une grande partie de ce chapitre a été consacrée aux détails d'implémentation de notre approche, ainsi qu'à la présentation et l'explication des résultats obtenus.

Une interprétation biologique assez large de ces résultats, renforcée par des tests de validation (via la plateforme String-DB.org) des co-expressions découvertes, ont été fournies. Un comparatif entre les résultats des deux modélisations (floue et possibiliste) montre que l'approche floue a détecté des relations fortes entre des sous-ensembles de gènes, mais que la modélisation possibiliste était plus fidèle à la réalité biologique des images étudiées, en conservant les détails in situ des zones d'expressions génétiques.

Conclusion Générale

L'objectif de ce travail était de modéliser l'aspect incertain des données d'Exp_Gen, afin d'extraire des RA floues et possibilistes spatio-temporelles, avec pour but de fournir une description formelle de la relation cachée entre les gènes qui co-expriment dans les séquences d'images ISH de l'espèce modèle "Mouse Edinburg", qui est une source riche de connaissances cruciales sur le développement de l'embryon des espèces vivantes.

Initialement, une série d'opérations de prétraitement a été appliquée pour extraire les caractéristiques les plus pertinentes de ces images, selon plusieurs critères : niveau d'Exp_Gen, variabilité et α-coupe du *Deg_App* dans la modélisation floue et possibiliste. Ainsi, deux tables de transactions floue et possibiliste ont été extraites, conservant la quasi-totalité des détails (forces et zones) des expressions génétiques dans les images du dataset Emage, avec une réduction claire de la dimensionnalité de ce dataset. L'adaptation proposée de l'algorithme Apriori (initialement conçu pour les données binaires) aux données incertaines d'Exp_Gen a aidé à extraire une variété d'itemsets fréquents (flous et possibilistes), ainsi que des RA utiles dont la pertinence a été prouvée via divers métriques et outils de validation solides. Cette adaptation de l'Apriori était basée sur une redéfinition des concepts de l'*union* et de la *cardinalité* des sous-ensembles. Notre processus, appliqué en Primal et en Dual, a découvert trois types de relations biologiques entre : gènes, organes (zones du corps), et phases de développement de l'espèce modèle étudiée.

La comparaison entre les résultats de ces deux modélisations montre qu'ils découvrent des RA d'un grand intérêt biologique, en particulier l'approche possibiliste, qui s'avère la plus adéquate pour modéliser l'incertitude dans les données étudiées. La méthode MAP préserve mieux les limites entre les zones d'Exp_Gen chevauchées, et fournit une résistance aux pixels bruités. De plus, cette modélisation favorise les RA basées sur des cases ayant subi une Exp_Gen claire, ce qui améliorera la qualité des connaissances extraites et leur pertinence biologique.

La modélisation du cas normal d'Exp_Gen pendant les phases de développement des espèces vivantes aidera à comprendre les causes des maladies génétiques suite à des expressions anormales. Pour améliorer ce travail, nous suggérons comme perspectives : une modélisation basée sur la logique évidentielle (théorie de Dempster-Shafer) en redéfinissant l'union et la cardinalité des itemsets suivant les fonctions de masse de croyance. De plus, l'exploration des motifs séquentiels fréquents (via l'algorithme AprioriAll) peut aider à extraire des relations temporelles entre les gènes pendant les différentes phases de développement des embryons.

Annexe : Contributions scientifiques

Conférence : Les 1ères Journées Scientifiques sur l'Informatique et ses Applications, *Universite 8 Mai 1945 de Guelma, Algérie,* 03-04 Mars 2009.

Titre : Intégration des techniques du Datamining dans le processus de Gestion des Connaissances basée sur le raisonnement à partir de cas

Auteurs : N. Mekroud, A. Moussaoui, Y. Slimani

Conférence : Conférence internationale sur l'informatique et ses applications, *Université de Saida, Algérie,* 03-04 Mai 2009.

Titre : Approche hybride pour le diagnostic industriel basée sur le RàPC et le Datamining : Utilisation de la plateforme JCOLIBRI 2.1

Auteurs : N. Mekroud, A. Moussaoui

Conférence : Conférence Internationale des Technologies de l'Information et de la Communication (CITIC'09), *Université de Setif-1, Algérie*, 04-05 Mai 2009.

Titre : Intégration des techniques du Datamining et des bases de données avancées dans le processus de Gestion des Connaissances : proposition d'un processus hybride basé sur le raisonnement à partir de cas.

Auteurs : N. Mekroud, A. Moussaoui, Y. Slimani

Conférence : 4ème édition Atelier des Systèmes Décisionnels (ASD'09), *Université de Jijel, Algérie,* 10-11 Novembre 2009.

Titre : Capitalisation des connaissances pour le diagnostic industriel : approche hybride Datamining-RàPC

Auteurs : N. Mekroud, A. Moussaoui

Conférence : Second International Conference and School on Radiation Imaging and Nuclear Medicine. *Ferhat Abbas-Setif1 University-UFAS1, In partnership with the Atomic Energy Commission-COMENA, Setif, Algeria,* June 11-15, 2023.

Titre : Knowledge extraction from gene expression images: Adaptation of association rules mining for theory of evidence.

Auteurs : N. Mekroud, A. Moussaoui

BIBLIOGRAPHIE

[1] T. Denecker. "Bioinformatique et analyse de données multiomiques: principes et applications chez les levures pathogènes Candida glabrata et Candida albicans". *Thèse de doctorat*. Université Paris-Saclay, 2020.

[2] G. Deléage, M. Gouy. "Bioinformatique: cours et cas pratique". *Edition Dunod*, 2013.

[3] A. Mihi. "Détection d'événements par les méthodes intelligentes dans les séquences biomoléculaires". *Thèse de doctorat*. Université Ferhat ABBAS, Sétif-1. 2019.

[4] M. Zekri. "Approches Bio-inspirées pour la Fouille de Données en Bioinformatique". *Thèse de doctorat*. Université BADJI Mokhtar-Annaba, 2015.

[5] Y.S. Taleb, P. Memon, A. Jalbani, N. Al-Anazi, A. Al-Garni, M. Altaweel, ... & Z. Iqbal. "Genotype–phenotype correlations in inherited cardiomyopathies, their role in clinical decision-making, and implications in personalized cardiac medicine in multi-omics as well as disease modeling eras". *Saudi Journal for Health Sciences*, 14(1), 30-41, 2025.

[6] E. Jaspard. "Support cours bioinformatique et Biochimie". *Univ Angers*. https://biochimej.univ-angers.fr/Page2/bioinformatique/7ModuleBioInfoJMGE/4IntroDefBioInfo/1IntroDefBioInfo.htm. Dernier accés : Janvier 2026

[7] N. Férey, G. Bouyer, C. Martin, A. Drif, P. Bourdot, M. Ammi, ... & L. Autin. "Docking de protéines en réalité virtuelle. Une approche hybride et multimodale." *Revue des Sciences et Technologies de l'Information-Série TSI: Technique et Science Informatiques*. vol. 28, no 8, p. 983-1015. 2009.

[8] I. Quinkal. "Quelques termes-clef de biologie moléculaire et leur définition." *INRIA Rhône-Alpes*, 2003.

[9] N. Sad Houari. "Bioinformatique et modélisation". *Polycopie, Université des Sciences et de la Technologie*, Oran. 2020.

[10] L. Richardson, S. Venkataraman, P. Stevenson, Y. Yang, J. Moss, L. Graham, ... & C. Armit. "EMAGE mouse embryo spatial gene expression database: 2014 update". *Nucleic acids research*, vol. 42, no D1, p. D835-D844, 2014.

[11] L. Ruzicka, D. G. Howe, S. Ramachandran, S. Toro, C. E. Van Slyke, Y. M. Bradford, ... & M. Westerfield. "The Zebrafish Information Network: new support for non-coding genes, richer Gene Ontology annotations ". *Nucleic acids research*, vol. 47, no D1, p. D867-D873, 2019.

[12] J. Thurmond, J. L. Goodman, V. B. Strelets, H. Attrill, L. S. Gramates, ... & B. R. Calvi. "FlyBase 2.0: the next generation". *Nucleic acids research*, vol.47, noD1, p. D759-D765, 2019.

[13] EMAGE gene expression database (http://www.emouseatlas.org/emage/). Dernier accés : Janvier 2026

[14] K. Puniyani, E.P. Xing. "GINI: from ISH Images to Gene Interaction Networks". *PLoS computational biology*, *9*(10), e1003227, 2013.

[15] S. Liu, C. Xu, Y. Zhang, J. Liu, B. Yu, X. Liu, M. Dehmer. "Feature selection of gene expression data for cancer classification using double RBF-kernels." *BMC bioinformatics.* vol. 19, no 1, p. 396. 2018.

[16] D. Yao, J. Yang, X. Zhan, X. Zhan, Z. Xie. "A novel random forests-based feature selection method for microarray expression data analysis." *International journal of data mining and bioinformatics*. vol. 13, no 1, p. 84-101. 2015.

[17] M. Xi, J. Sun, L. Liu, F. Fan, X. Wu. "Cancer Feature Selection and Classification Using a Binary Quantum-Behaved Particle Swarm Optimization and Support Vector Machine." *Computational and mathematical Methods in Medicine. vol. 2016, no 1, p. 3572705.* 2016.

[18] D. Pradhan, S. Padhy, B. Sahoo. "Enzyme classification using multiclass support vector machine and feature subset selection." *Computational biology and chemistry*. , vol. 70, p. 211-219. 2017.

[19] Y. Zhang, Q. Deng, W. Liang, X. Zou. "An efficient feature selection strategy based on multiple support vector machine technology with gene expression data." *BioMed research international.* vol 2018, n°1, p 7538204. 2018.

[20] Z. Li, W. Xie, T. Liu. "Efficient feature selection and classification for microarray data." *PloS one. vol. 13, n° 8, p. e0202167.* 2018.

[21] P. Das, A. Roychowdhury, S. Das, S. Roychoudhury, S. Tripathy. "sigFeature: novel significant feature selection method for classification of gene expression data using support vector machine and t statistic". *Frontiers in genetics*. vol. 11, p. 247. 2020.

[22] S. Acharya, S. Saha, P. Pradhan. "Novel symmetry-based gene-gene dissimilarity measures utilizing Gene Ontology: Application in gene clustering". *Gene*, vol. 679, p. 341-351, 2018.
[23] N. Novoselova, I. Tom. "Entropy-based cluster validation and estimation of the number of clusters in gene expression data." *Journal of bioinformatics and computational biology.* vol. 10, no 05, p. 1250011. 2012.
[24] L. Sun, J. Xu, J. Yin. "An effective fuzzy kernel clustering analysis approach for gene expression data." *Bio-Medical Materials and Engineering*. vol. 26, no 1_suppl, p. S1863-S1869. 2015.
[25] S. Wenric, R. Shemirani. "Using supervised learning methods for gene selection in RNA-Seq case-control studies." *Frontiers in genetics*. vol. 9, p. 364621. 2018.
[26] N.T. Johnson, A. Dhroso, K.J. Hughes, D. Korkin. "Biological classification with RNA-seq data: Can alternatively spliced transcript expression enhance machine learning classifiers?." *Rna* . vol. 24, no 9, p. 1119-1132. 2018.
[27] D. Urda, J. Montes-Torres, F. Moreno, L. Franco, J.M. Jerez. "Deep learning to analyze RNA-seq gene expression data." *International work-conference on artificial neural networks*. Cham: Springer International Publishing, 2017.
[28] N.E.M. Khalifa, M.H.N. Taha, D.E. Ali, A. Slowik, A.E. Hassanien. "Artificial intelligence technique for gene expression by tumor RNA-Seq data: a novel optimized deep learning approach". *IEEE Access*, vol. 8, p. 22874-22883, 2020.
[29] X. Li, K. Wang, Y. Lyu, H. Pan, J. Zhang, D. Stambolian, K. Susztak, M.P. Reilly, G. Hu, M. Li. "Deep learning enables accurate clustering with batch effect removal in single-cell RNA-seq analysis." *Nature communications*. vol. 11, no 1, p. 2338. 2020.
[30] C. Arisdakessian, O. Poirion, B. Yunits, X. Zhu, L.X. Garmire. "DeepImpute: an accurate, fast, and scalable deep neural network method to impute single-cell RNA-seq data." *Genome biology*. vol. 20, no 1, p. 211. 2019.
[31] H. Peng, F. Long, J. Zhou, G. Leung, M.B. Eisen, E.W. Myers. "Automatic image analysis for gene expression patterns of fly embryos." *BMC cell biology. vol. 8, no Suppl 1, p. S7.* 2007.
[32] D.L. Mace, N. Varnado, W. Zhang, E. Frise, U. Ohler. "Extraction and comparison of gene expression patterns from 2D RNA in situ hybridization images." *Bioinformatics*. vol. 26, no 6, p. 761-769. 2010.
[33] X. Cai, H. Wang, H. Huang, C. Ding. "Joint stage recognition and anatomical annotation of drosophila gene expression patterns." *Bioinformatics*. , vol. 28, no 12, p. i16-i24. 2012.
[34] T. Zeng, R. Li, R. Mukkamala, J. Ye, S. Ji. "Deep convolutional neural networks for annotating gene expression patterns in the mouse brain". *BMC bioinformatics*, vol. 16, p. 1-10, 2015.
[35] J.W. Bohland, H. Bokil, S.D. Pathak, C.K. Lee, L. Ng, C. Lau, C. Kuan, M. Hawrylycz, P.P. Mitra. "Clustering of spatial gene expression patterns in the mouse brain and comparison with classical neuroanatomy". *Methods*, vol. 50, no 2, p. 105-112, 2010.
[36] Y. Li, H. Huang, H. Chen, T. Liu. "Deep neural networks for in situ hybridization grid completion and clustering". *IEEE/ACM transactions on computational biology and bioinformatics*. vol. 17, no 2, p. 536-546. 2018.
[37] A. Andonian, D. Paseltiner, T.J. Gould, J.B. Castro. "A deep learning based method for large-scale classification, registration, and clustering of in-situ hybridization experiments in the mouse olfactory bulb". *Journal of neuroscience methods. vol. 312, p. 162-168.* 2019.
[38] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth. "From data mining to knowledge discovery in databases". *AI magazine*, vol. 17, no 3, p. 37-37. 1996.
[39] D.A. Zighed, R. Rakotomalala. "Extraction de connaissances à partir de données (ECD) ". *Techniques de l'Ingénieur. H*, vol. 3, p. 744. 2002.
[40] I. Rasovska. "Contribution à une méthodologie de capitalisation des connaissances basée sur le raisonnement à partir de cas: Application au diagnostic dans une plateforme d'e-maintenance". *Doctoral dissertation*, Université de Franche-Comté. 2006.
[41] C.A. Azencott. "Introduction au machine learning". *Malakoff*, France: Dunod. 2018.
[42] S.J. Russell, P. Norvig. "Artificial Intelligence : A Modern Approach". *Global Edition 4e*. 2021.
[43] R. Agrawal, T. Imieliński, A. Swami. "Mining association rules between sets of items in large databases". In : *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*. p. 207-216. 1993.
[44] P. Hájek, I. Havel, M. Chytil. "The GUHA method of automatic hypotheses determination". *Computing, vol. 1, no 4, p. 293-308*. 1966.

[45] J.L. Guigues, V. Duquenne. "Familles minimales d'implications informatives résultant d'un tableau de données binaires". *Mathématiques et Sciences humaines*, *vol. 95, p. 5-18*. 1986.
[46] D. Gangaramani, R. Londhe. "Survey on association rule analysis: Exploration using mining analysis". *International Journal of Hybrid Intelligent Systems*, vol. 21, no 1, p. 2-13. 2025.
[47] J. Hipp, U. Güntzer, G. Nakhaeizadeh. "Algorithms for association rule mining - a general survey and comparison". *ACM sigkdd explorations newsletter*, vol. 2, no 1, p. 58-64. 2000.
[48] M.L. Anne, M.T. Maguelonne, M.A.M. Rachid, M.B. Tassadit, M.E.M. Sami, M.K. Mamadou. "Fouille de données : Règles séquentielles". *Université Montpellier II*, 2008.
[49] S. Lallich, O. Teytaud. "Évaluation et validation de l'intérêt des regles d'association". *Revue des nouvelles Technologies de l'Information*, vol. 1, no 2, p. 193-218. 2004.
[50] B. Vaillant, P. Meyer, E. Prudhomme, S. Lallich, P. Lenca, S. Bigaret. "Mesurer l'intérêt des règles d'association". *In : Atelier Qualité des Données et des Connaissances (DKQ 2005) associé à EGC*. pp. 421-426. 2005.
[51] N. Lavrač, P. Flach, B. Zupan. "Rule evaluation measures: A unifying view". *In : International Conference on Inductive Logic Programming*. Berlin, Heidelberg : Springer Berlin Heidelberg, p. 174-185. 1999.
[52] P.D. Mcnicholas. "Standardising the lift of an association rule". *Computational Statistics & Data Analysis*, vol. 52, no 10, p. 4712-4721. 2008.
[53] G. Piatetsky-Shapiro. "Discovery, Analysis, and Presentation of Strong Rules". *Knowledge Discovery in Databases*. AAAI/MIT Press, Cambridge, 248, 255-264. 1991.
[54] R. Agrawal, R. Srikant. "Fast algorithms for mining association rules". *In : 20th VLDB Conference*. p. 487-499. 1994.
[55] A. Savasere, E. Omiecinski, S. Navathe. "An Efficient Algorithm for Mining Association Rules in Large Databases". *In : Proceedings of the 21st International Conference on Very Large Databases (VLDB)*. p. 432-444. 1995.
[56] M.J. Zaki, S. Parthasarathy, M. Ogihara, W. Li. "New algorithms for fast discovery of association rules". *In : KDD*. Vol. 97, p. 283-286. 1997.
[57] J. Han, J. Pei, Y. Yin. "Mining frequent patterns without candidate generation". *ACM sigmod record*, vol. 29, no 2, p. 1-12. 2000.
[58] M. Kantardzic. "Data Mining: Concepts, models, methods, and algorithms". *Technometrics*, vol. 45, no 3, p. 277. 2003
[59] B. Brühl, M. Hülsmann, D. Borscheid, C.M. Friedrich, D. Reith. "A sales forecast model for the german automobile market based on time series analysis and data mining methods". *In : Industrial Conference on Data Mining*. Berlin, Heidelberg : Springer Berlin Heidelberg. p. 146-160. 2009.
[60] T. Piton, J. Blanchard, F. Guillet, H. Briand. "Une méthodologie de recommandations produits fondée sur l'actionnabilité et l'intérêt économique des clients". *In : Extraction et gestion des connaissances (EGC'2011)*. Hermann, p. 203-214. 2011.
[61] R. Srikant, R. Agrawal. "Mining quantitative association rules in large relational tables". In : *Proceedings of the 1996 ACM SIGMOD international conference on Management of data*. p. 1-12, 1996.
[62] R. Srikant, R. Agrawal. "Mining generalized association rules". *Future generation computer systems*, vol. 13, no 2-3, p. 161-180. 1997.
[63] R. Agrawal, R. Srikant. "Mining sequential patterns". In : *Proceedings of the eleventh international conference on data engineering*, IEEE, pp. 3–14. 1995.
[64] L.A. Zadeh. "Fuzzy sets". *Information and Control*, vol. 8, no 3, p. 338-353. 1965.
[65] Schéma de l'ADN et chromosomes (https://www.futura-sciences.com/sante/definitions/medecine-adn-87/). Dernier accés : Janvier 2026
[66] V. Panneton. "Régulation de la kinase Ste20 Slik de Drosophila par phosphorylation de la boucle d'activation". *Programme de Biologie Moléculaire Faculté de Médecine*. Univ de Montréal. 2014.
[67] C. M. Kuok, A. Fu, and M. H. Wong, "Mining fuzzy association rules in databases". *SIGMOD Rec.*, vol. 27, no. 1, pp. 41–46. 1998.
[68] B. Bouaita, A. Moussaoui, N.I. Bachari. "Rainfall estimation from MSG images using fuzzy association rules". *Journal of Intelligent & Fuzzy Systems*, vol. 37, no 1, p. 1357-1369. 2019.
[69] J.M. de Graaf, W.A. Kosters, J.J.W. Witteman. "Interesting fuzzy association rules in quantitative databases". In : *European Conference on Principles of Data Mining and Knowledge Discovery*. Berlin, Heidelberg: Springer Berlin Heidelberg, p. 140-151. 2001.

[70] C.H. Weng, Y.L. Chen. "Mining fuzzy association rules from uncertain data". *Knowledge and Information Systems*, vol. 23, no 2, p. 129-152. 2010.
[71] L.A. Zadeh. "Fuzzy sets as a basis for a theory of possibility". *Fuzzy Sets and Systems*, vol. 1, no 1, p. 3-28. 1978.
[72] Y. Djouadi, S. Redaoui, K. Amroun. "Mining association rules under imprecision and vagueness: towards a possibilistic approach". In : *2007 IEEE International Fuzzy Systems Conference*. IEEE, p. 1-6, 2007.
[73] A.P. Dempster. "Upper and lower probability inferences based on a sample from a finite univariate population". *Biometrika*, vol. 54, no 3-4, p. 515-528. 1967.
[74] G. Shafer. "A Mathematical theory of Evidence". *Princeton University Press*, 1976.
[75] M.A.B. Tobji, B.B. Yaghlane, K. Mellouli. "A new algorithm for mining frequent itemsets from evidential databases". In : *Proceedings of IPMU*, p. 1535-1542. 2008.
[76] J.P. Lavergne. Socrate, platon et aristote : Les débuts de la 'sagesse'. Vie des Sciences, Sociétés, Arts. https://www.jeanpierrevarlenge.com/. Dernier accés : Janvier 2026
[77] G. Boole. "An Investigation of the Laws of Thought, on Which Are Founded the Mathematical Theories of Logic and Probabilities". *London: Macmillan*. 1854.
[78] M. Serfati. "A la recherche des Lois de la Pensée. Sur l'épistémologie du calcul logique et du calcul des probabilités chez Boole". *Mathématiques et sciences humaines*. Mathematics and social sciences, no 150. 2000.
[79] M-H Masson. "Apports de la théorie des possibilités et des fonctions de croyance à l'analyse de données imprécises". *Habilitation à diriger des Recherches*, vol. 2. 2005.
[80] J.M. Tacnet, M. Batton-Hubert, J. Dezert. "Analyse multicritères et fusion d'information pour l'expertise et la gestion intégrée des risques naturels en montagne". In : *Colloque LambdaMu 17*. p. 10 p. 2010.
[81] C. De Runz. "Imperfection, temps et espace : modélisation, analyse et visualisation dans un SIG archéologique". *Thèse de doctorat*. Université de Reims-Champagne Ardenne. 2008.
[82] M. Ekovich. "Contrôle exécutif et mémoire : inférences probabilistes et récupérations mnésiques dans les lobes frontaux". *Thèse de doctorat*. Université Pierre et Marie Curie-Paris VI. 2014.
[83] P.A. Bisgambiglia. "Approche de modélisation approximative pour des systèmes à événements discrets: Application à l'étude de propagation de feux de forêt". *Thèse de doctorat*. Université Pascal Paoli. 2008.
[84] D. Dubois, H. Prade. "Possibility theory: an approach to computerized processing of uncertainty". *Plenum Press*, NY, 1988.
[85] B. Elayeb. "Recherche d'Information Possibiliste : De la Désambiguïsation et la Reformulation de Requêtes vers la Fiabilité de l'Information Recherchée". *Habilitation Universitaire*. Manouba University; National School of Computer Science (ENSI). 2018.
[86] D. Dubois, H. Prade. "Théorie des possibilités". *Institut de Recherche en Informatique de Toulouse*. 2006.
[87] A. Samet. "Théorie des fonctions de croyance : application des outils de datamining pour le traitement de données imparfaites". *Thèse de doctorat*. UTM-FST (Tunisie) - UA-LGI2A (France). 2014.
[88] D. Mercier. "Concepts de base de la théorie des fonctions de croyance". https://www.davidmercier.fr/teaching/cours_fc.pdf. Dernier accès : janvier 2026
[89] T. Denoeux, D. Dubois, H. Prade. "Representations of Uncertainty in Artificial Intelligence: Probability and Possibility". *A Guided Tour of AI Research Volume I.* Springer International Publishing, pp.69-117, 10.1007/978-3-030-06164-7_3. hal-02921346. 2020.
[90] D. Dubois, E. Hüllermeier, and H. Prade. "A systematic approach to the assessment of fuzzy association rules". *Data Mining and Knowledge Discovery*, vol. 13, no. 2, pp. 167–192. 2006.
[91] M. Delgado, N. Marín, D. Sánchez, M.A. Vila. "Fuzzy association rules: general model and applications". *IEEE transactions on Fuzzy Systems*, vol. 11, no 2, p. 214-225. 2003.
[92] T.P. Hong, K.Y. Lin, S.L. Wang. "Fuzzy data mining for interesting generalized association rules". *Fuzzy sets and systems*, vol. 138, no 2, p. 255-269. 2003.
[93] C. Fiot, G. Dray, A. Laurent, M. Teisseire. "A la recherche des motifs sequentiels flous". *Actes des Rencontres Francophones sur la Logique Floue et ses Applications (LFA 04)*, pages 131–138, 2004.
[94] J. C. Bezdek. "Objective function clustering". In : *Pattern recognition with fuzzy objective function algorithms*, pages 43–93. Springer, 1981.

[95] R. Pierrard, J.-P. Poli, and C. Hudelot. "A fuzzy close algorithm for mining fuzzy association rules". In : *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 88–99. Springer, 2018.
[96] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. "Pruning closed itemset lattices for association rules". *BDA'1998 international conference on Advanced Databases*, Hammamet, Tunisia. pp.177-196. ffhal-00467745f. 1998.
[97] T.-P. Hong, C.-S. Kuo, S.-L. Wang. "A fuzzy AprioriTid mining algorithm with reduced computational time". *Applied Soft Computing*, vol. 5, no 1, p. 1-10. 2004.
[98] G. Kamingu. "Théorie des ensembles flous : caractérisation, propriétés et opérations". *One Pager Lareq*, vol. 11, p. 37-45. 2016.
[99] S.C. Chen, T.H. Tsai, C.H. Chung, W.H. Li. "Dynamic association rules for gene expression data analysis". *BMC genomics*, vol. 16, no 1, p. 786. 2015.
[100] A. Houari, W. Ayadi, S. B. Yahia. "Mining negative correlation biclusters from gene expression data using generic association rules". *Procedia computer science*, vol. 112, p. 278-287, 2017.
[101] K. Vougas, M. Krochmal, T. Jackson, A. Polyzos, ... & V.G. Gorgoulis. "Deep learning and association rule mining for predicting drug response in cancer. A personalised medicine approach." *BioRxiv*, p. 070490. 2016.
[102] P. Cremaschi, R. Carriero, S. Astrologo, C. Colì, A. Lisa, S. Parolo, S. Bione. "An association rule mining approach to discover lncRNAs expression patterns in cancer datasets". *BioMed research international*, vol. 2015, no 1, p. 146250. 2015.
[103] R. Vengateshkumar, S. Alagukumar, R. Lawrance. "Boolean Association Rule Mining on Microarray Cancer Gene Expression Data using Gene Expression Intervals". In : *International Journal of Innovative Research in Computer and Communication Engineering*. 2017.
[104] R. Vengateshkumar, S. Alagukumar, R. Lawrance. "Boolean association rule mining on microarray gene expression data". In : *Advanced Computing and Intelligent Engineering: Proceedings of ICACIE 2018, Volume 1*. Springer Singapore, p. 99-111, 2020.
[105] J. Van Hemert, R. Baldock. "Mining spatial gene expression data for association rules". In : *International Conference on Bioinformatics Research and Development*. Springer: Berlin Heidelberg, p. 66-76. 2007.
[106] M. Anandhavalli, M.K. Ghose, K. Gauthaman. "Interestingness measure for mining spatial gene expression data using association rule". *arXiv preprint arXiv:1001.3498.* 2010.
[107] Y. He. S. C. Hui. "Exploring Ant-Based Algorithms for Gene Expression Data Analysis". *Artificial Intelligence in Medicine*, vol. 47, no 2, p. 105-119. 2009.
[108] Y. Gheraibia, A. Moussaoui, Y. Djenouri, S. Kabir, P.Y. Yin. "Penguins search optimisation algorithm for association rules mining". *Journal of computing and information technology,* vol. 24, no 2, p. 165-179. 2016.
[109] P. Das, R. Islam, N. Saha, S. Mitra, S. Chandra. "Association Rule Mining for Gene Expression Data -A Neural Network Approach". *International Journal of Engineering Trends and Applications (I JETA)* – Volume 3 Issue 2. 2016.
[110] S. Alagukumar, C.D.A. Vanitha, R. Lawrance. "Clustering of Association Rules on Microarray Gene Expression Data". In : *Advanced Computing and Intelligent Engineering: Proceedings of ICACIE 2018, Volume 1*. Singapore : Springer Singapore, p. 85-97. 2020.
[111] P. Sethi, S. Alagiriswamy. "Association rule based similarity measures for the clustering of gene expression data". *The open medical informatics journal*, vol. 4, p. 63. 2010.
[112] A. Hamdi, K. Shaban, A. Erradi, A. Mohamed, S.K. Rumi, F.D. Salim. "Spatiotemporal data mining: a survey on challenges and open problems". *Artificial Intelligence Review*, vol. 55, no 2, p. 1441-1488. 2022.
[113] K. Kianmehr, M. Kaya, A.M. ElSheikh, J. Jida, R. Alhajj. "Fuzzy association rule mining framework and its application to effective fuzzy associative classification". *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 6, pp. 477–495, 2011.
[114] X. Geng, Y. Liang, L. Jiao. "EARC: Evidential association rule-based classification". *Information Sciences*, vol. 547, pp. 202–222. 2021.
[115] S. H. Bouazza. "A Deep Ensemble Gene Selection and Attention-guided Classification Framework for Robust Cancer Diagnosis from Microarray Data". *Engineering, Technology & Applied Science Research*, vol. 15, no. 1, pp. 20235–20241. 2025.

[116]M. Ciortan, M. Defrance. "GNN-based embedding for clustering scRNA-seq data". *Bioinformatics*, vol. 38, no. 4, pp. 1037–1044. 2022.
[117]S. Papadimitriou, S. Mavroudi, S.D. Likothanassis. "Mutual information clustering for efficient mining of fuzzy association rules with application to gene expression data analysis". *International Journal on Artificial Intelligence Tools*, vol. 15, no. 02, pp. 227–250. 2006.
[118]M. Pereyra. "Maximum-A-Posteriori estimation with Bayesian confidence regions". *SIAM Journal on Imaging Sciences*, vol. 10, no 1, p. 285-302. 2017.
[119]D. Szklarczyk, R. Kirsch, M. Koutrouli, K. Nastou, F. Mehryary, R. Hachilif, ... & C. Von Mering. "The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest". *Nucleic Acids Research*. 2023 Jan 6;51(D1):D638-646. URL : https://string-db.org/