Intégration de motifs discriminants dans un graphe de connaissances pour la prédiction de la forme de la Leucémie Lymphoïde Chronique à partir des séquences de valeurs du MTS et des données patients

Amal Beldi[1], Nathalie Pernelle[1], Sylvie Després[2], Rahma Dandan[2], Claudine Irles [3], Christine Le Roy[3]

[1] Université Sorbonne Paris Nord, UMR 7030, LIPN
[2] Université Sorbonne Paris Nord, UMRS 1142, LIMICS
[3] Université Sorbonne Paris Nord, INSERM UMR 1349, SIMHEL

Résumé

Certains patients atteints de leucémie lymphoïde chronique (LLC) restent cliniquement stables, et d'autres présentent une évolution rapide nécessitant un traitement. Nous proposons une chaîne de traitement originale visant à évaluer la capacité prédictive de motifs extraits des trajectoires longitudinales des valeurs normalisées du MTS ou de leurs vitesses. De plus, un graphe de connaissances, enrichi avec les motifs discriminants, est exploité afin de découvrir des règles associées à la forme de la LLC au sein des sous-populations de patients. Les expériences menées sur des données réelles montrent que ces deux hypothèses semblent vérifiées.

Mots-clés

Stratification des données longitudinales, Graphe temporel de connaissances, Prédiction d'évolution de la leucémie lymphoïde chronique.

Abstract

Some patients with chronic lymphocytic leukemia (CLL) remain clinically stable, while others exhibit rapid progression requiring early therapeutic intervention. We propose an original processing pipeline that allows biologists to verify that CLL subtype prediction can be based either on patterns representing the dynamics of MTS time series or on their normalized values. Furthermore, a knowledge graph enriched with discriminative patterns is leveraged to discover rules that infer the CLL form for patient subpopulations. Experiments conducted on real data indicate that both of these hypotheses appear to be validated.

Keywords

Longitudinal data stratification, Temporal Knowledge Graph, Prediction of chronic lymphocytic leukemia progression.

1 Introduction

La leucémie lymphoïde chronique (LLC) est un cancer hématologique qui survient lorsque certains globules blancs, les lymphocytes B, acquièrent une durée de vie anormalement longue. Il s'agit de la forme la plus fréquente de leucémie chez l'adulte. La LLC présente une forte hétérogénéité clinique, avec des trajectoires d'évolution de la maladie très variables d'un patient à l'autre. Alors que certains patients restent indolents pendant de nombreuses années, d'autres développent une progression rapide nécessitant une prise en charge thérapeutique. Traditionnellement, l'évaluation du risque évolutif de la LLC repose sur des caractéristiques cliniques du patient et sur différents facteurs de risque biomarqueurs [1] comme les mutations somatiques des gènes des chaînes lourdes des immunoglobulines (IGHV) et les anomalies cytogénétiques. En revanche, aucun de ces facteurs pris isolément ne possède une valeur prédictive suffisamment robuste à l'échelle individuelle pour anticiper la progression de la maladie. Afin de proposer une alternative à ces facteurs cliniques et biologiques, le laboratoire SIMHEL utilise les résultats du test MTS inspiré du test de viabilité cellulaire MTT, pour mesurer l'activité métabolique des lymphocytes B ex-vivo suite à une stimulation antigénique. Les travaux expérimentaux du SIMHEL montrent que le MTS a une valeur prédictive en termes de survie globale et de progression des patients LLC [3]. Enfin, l'hypothèse est également faite que les motifs discriminants découverts pourraient avoir une capacité prédictive différente selon les caractéristiques du patient (e.g., sexe, caractéristiques cytogénétiques).
Cette étude a pour objectif d'évaluer la valeur prédictive des profils évolutifs des patients en fonction des valeurs MTS. La série temporelle des valeurs de ce biomarqueur potentiel pourrait être analysée non pas seulement en termes d'amplitudes normalisées ou de seuils de valeur atteints, mais aussi en termes de variations relatives. Malgré la richesse des données longitudinales disponibles, plusieurs défis limitent leur exploitation en pratique : (1) les mesures de MTS ne sont pas acquises aux mêmes intervalles de temps pour tous les patients, rendant les comparaisons directes difficiles, (2) les séquences complètes de valeurs numériques de vitesse sont difficiles à interpréter, (3) les cohortes étudiées restent de taille limitée. L'objectif est par conséquent de transformer des données longitudinales hétérogènes de MTS en représentations transformées, interprétables et discrimi-

1. Un biomarqueur est une caractéristique mesurable renseignant sur une fonction biologique, un processus pathologique ou une réponse biologique à un traitement thérapeutique [2]

nantes permettant de caractériser les profils évolutifs des patients atteints de LLC. Nous proposons une chaîne de traitements originale permettant d'évaluer la valeur prédictive du MTS. La classification des patients peut s'effectuer soit sur les motifs représentant la dynamique des vitesses du MTS (i.e. motifs symboliques ou numériques), soit sur les séquences de valeurs normalisées harmonisées temporellement. De plus, un graphe de connaissances nommé KG_{LLC} dédié à la représentation des connaissances sur les patients, enrichi avec les motifs discriminants, est exploité afin de découvrir des règles associant certaines caractéristiques à la forme de la LLC au sein de sous-populations de patients. Les expériences menées sur un jeu de données réel de 102 patients montrent que des motifs discriminants peuvent être identifiés, et que la valeur prédictive de ce biomarqueur potentiel varie en fonction des caractéristiques des patients pris en compte.
Dans cet article, nous proposons les contributions suivantes :

— Une transformation de séries longitudinales irrégulières en représentations comparables fondées sur les valeurs et les vitesses du MTS.

— Une extraction conjointe de motif discriminant symbolique (PrefixSpan) et numérique (Shapelets).

— Une intégration de ces motifs dans un graphe de connaissances pour permettre l'apprentissage de règles interprétables.

— Des expérimentations montrant que la dynamique du MTS possède un pouvoir prédictif supérieur à celui des valeurs normalisées.

2 Contexte et problématique

2.1 Facteurs pronostiques de la LLC

Les facteurs pronostiques de la LLC sont des éléments qui permettent d'associer un risque à la progression de la maladie pour prendre des décisions thérapeutiques en tenant compte du rapport bénéfice/risque. Il est important de disposer de facteurs fiables qui prennent en compte la variabilité clinique de la maladie d'un patient à l'autre. Parmi les facteurs ayant un fort impact sur le pronostic figurent : le temps de doublement des lymphocytes sur une période de 3-6 mois, un statut non muté IGHV, la présence de délétions chromosomiques telles que la del17q ou del11q. Notre approche se concentre sur la prédiction de la forme indolente et progressive chez les patients LLC. Parmi les mesures longitudinales disponibles, le valeurs de MTS constitueraient un marqueur potentiel de la dynamique évolutive de la LLC.

2.2 Problématique

Un patient est caractérisé par une combinaison de descripteurs statiques et de données longitudinales collectées au cours du suivi clinique. Les patients sont différenciés par leur appartenance à la classe des patients progressifs (P) ou indolents (I) (i.e. non progressif). En outre, chaque patient p est associé à une séquence $\mathcal{O}_p$ de valeurs de MTS selon des intervalles de temps dépendants de son état :

$$\mathcal{O}_p = \{(t_1, m_1), (t_2, m_2), \ldots, (t_{n_p}, m_{n_p})\},$$

où $t_{p,i} \in \mathbb{R}^+$ dénote le nombre de mois écoulés depuis la première collecte des données du patient, $m_{p,i} \in \mathbb{R}$ représente la valeur de MTS (en pourcentage), et n_p est le nombre d'observations disponibles pour le patient p.
De plus, chaque patient est associé à un vecteur de descripteurs statiques,

$$\mathbf{x}_p = (x_p^{\text{age}}, x_p^{\text{sex}}, x_p^{\text{diag}}, x_p^{\text{treat}}, x_p^{\text{gen}}, \ldots),$$

Ce vecteur comprend les informations démographiques (e.g. l'âge et le sexe), les attributs liés au diagnostic, l'historique des traitements, ainsi que les résultats issus d'analyses biologiques ou génétiques lorsqu'ils sont disponibles. Ces variables sont considérées comme invariantes dans le temps.

L'objectif principal est de découvrir des ensembles de motifs permettant de représenter des variations du MTS qui sont susceptibles de discriminer les formes de la LLC. Ces motifs peuvent être symboliques, ou présentés sous la forme de Shapelets (i.e. séquences numériques discriminantes associées à une distance maximale). Deux types de patterns sont considérés : les patterns issus d'une description de la dynamique des valeurs du MTS (e.g. croissance modérée, décroissance rapide, stabilité) et des patterns issus des séquences de valeurs collectées.

Nous proposons de construire un graphe temporel de connaissances, nommé KG_{LLC}. Ce choix afin de bénéficier d'une expressivité suffisante pour représenter explicitement les concepts, les relations et les contraintes sémantiques, ainsi que d'une compatibilité avec des outils d'apprentissage de règles. Le graphe intègre à la fois des données longitudinales ainsi que des propriétés dérivées telles que les motifs discriminants.

3 Etat de l'art

3.1 Classification de séries temporelles

Dans ce travail, nous nous concentrons sur des approches de classification de séries temporelles permettant l'extraction de motifs interprétables à partir de données longitudinales. Contrairement aux approches basées sur l'apprentissage profond, nous privilégions des méthodes explicables adaptées au contexte biomédical.
Méthodes fondées sur des motifs numériques (shapelets). Les approches basées sur les Shapelets identifient des sous-séquences locales discriminantes, caractéristiques de classes spécifiques [9, 4]. Ces méthodes permettent de capturer des motifs temporels localisés et offrent un bon compromis entre performance et interprétabilité. Toutefois, leur coût computationnel peut être élevé et elles sont sensibles au choix des hyperparamètres.
Méthodes symboliques. Certaines approches telles que SAX (Symbolic Aggregate approXimation) [8] discrétisent les séries temporelles pour les modéliser comme des collections de mots. Ces représentations peuvent constituer la base de classifieurs performants tels que WEASEL [14]. Les méthodes d'extraction de motifs séquentiels, comme PrefixSpan [11] opèrent directement sur des séquences symboliques

afin d'extraire des motifs temporels (non-)contigus fréquents qui peuvent être adaptés à la validation des biologistes.
Ces approches permettent d'extraire des motifs explicites directement exploitables sous forme de règles.

3.2 Graphes de connaissances temporels pour les données longitudinales en santé

Dans le domaine médical, les événements cliniques sont dynamiques et fortement dépendants du temps (progression de maladies, traitements, etc.). Les graphes de connaissances temporels (TKGs) permettent de représenter l'évolution temporelle de l'état de santé des patients et les événements cliniques sont représentés sous forme de triplets enrichis d'une dimension temporelle [5]. Des approches telles que MedTKG [13] combinent des graphes dynamiques et des connaissances médicales structurées. Toutefois, ces approches reposent généralement sur des représentations latentes complexes, ce qui limite leur interprétabilité dans un contexte clinique.

3.3 Apprentissage de règles dans les graphes de connaissances

L'apprentissage de règles dans les graphes de connaissances vise à extraire des règles de Horn en logique du premier ordre, permettant des inférences explicables.
AMIE3 [7] est une approche efficace d'extraction de règles basée sur une stratégie top-down, intégrant des techniques d'élagage et des mesures adaptées aux graphes incomplets telles que la confiance PCA. Des approches alternatives comme AnyBURL [10] reposent sur des marches aléatoires dans le graphe afin de générer des règles de manière incrémentale. Certaines méthodes étendent ces approches pour intégrer des contraintes numériques, comme REGNUM [6], qui enrichit les règles symboliques avec des intervalles de valeurs. Ces extensions peuvent être particulièrement pertinentes dans les contextes biomédicaux où les données numériques jouent un rôle central.

3.4 Positionnement

Notre approche combine l'extraction de motifs discriminants à partir de séries temporelles (shapelets et motifs séquentiels) avec l'apprentissage de règles dans un graphe de connaissances. Son originalité réside dans l'intégration de motifs issus des valeurs et des vitesses d'un biomarqueur longitudinal au sein d'une représentation sémantique unifiée.
Cette intégration permet de contextualiser les motifs en fonction des caractéristiques des patients et d'induire des règles explicables en logique du premier ordre. Contrairement aux approches basées sur des représentations latentes, notre méthode privilégie l'interprétabilité et la capacité de validation par des experts, ce qui constitue un enjeu essentiel dans un contexte biomédical.

4 Approche proposée

Nous proposons une chaîne de traitement visant à (i) transformer des données longitudinales hétérogènes en représentations comparables, (ii) extraire des motifs discriminants interprétables à partir de ces représentations, et (iii) intégrer ces motifs à un graphe de connaissances peuplé afin d'induire des règles de classification interprétables (cf. figure 1). L'approche repose sur une double représentation du biomarqueur MTS : les valeurs normalisées harmonisées temporellement, et les vitesses de variation à pas semestriel.

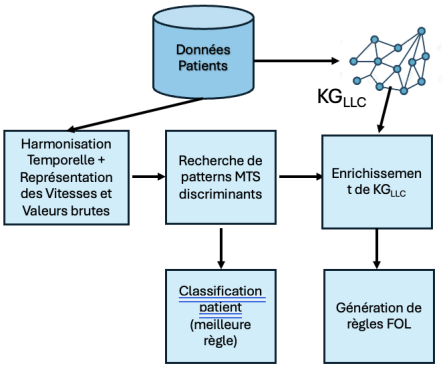


FIGURE 1 – Pipeline global de l'approche.

4.1 Harmonisation temporelle des séquences de valeurs de MTS

L'entrée du pipeline est constituée des trajectoires longitudinales du marqueur MTS pour chaque patient p :

$$O_p = \{(t_1, m_1), \ldots, (t_{n_p}, m_{n_p})\},$$

où t_i représente le temps (en mois) et m_i la valeur du MTS. Ces mesures étant acquises à intervalles irréguliers, une harmonisation temporelle est nécessaire. Les valeurs de MTS ont été reconstruites sur une grille temporelle T régulière à pas de 6 mois (i.e. T = {0, 6, 12, 18, ... }) à partir de t0 qui correspond à la collecte de la première valeur de MTS pour chaque patient, et ceci jusqu'au dernier temps après lequel une valeur est observée. Plus précisément, pour chaque temps $t_k \in T$:

— Si une mesure de MTS est observée exactement à t_k, la valeur mesurée est conservée.

— Sinon, si t_k est compris entre deux observations consécutives, i.e. $t_j < t_k < t_{j+1}$, une interpolation linéaire est appliquée. La valeur interpolée $\widehat{m}(t_k)$ est donnée par :

$$\widehat{m}(t_k) = m_j + \frac{m_{(j+1)} - m_j}{t_{(j+1)} - t_j}\,(t_k - t_j)\,.$$

En conséquence, chaque patient est associé à un vecteur de séquence de valeurs brutes temporellement normalisées $\mathcal{V}al = (val_1, val2, \ldots, val_{n_p})$. Cette approche fait l'hypothèse qu'une estimation raisonnable peut être faite en supposant une évolution linéaire locale du marqueur étudié entre deux observations successives. Aucune extrapolation n'est réalisée en dehors de l'intervalle des observations disponibles. Si t_k est postérieur à la dernière mesure observée, la valeur est considérée comme manquante.

Ce procédé permet d'obtenir, pour chaque patient, une trajectoire MTS alignée temporellement, facilitant les analyses longitudinales comparatives et les approches supervisées d'apprentissage de motifs discriminants.

4.2 Représentation des séquences de vitesses du MTS

En complément des trajectoires brutes de MTS harmonisées tous les six mois, nous avons construit des trajectoires permettant de représenter la façon dont le MTS évolue au cours du temps afin de capturer la dynamique d'évolution du score (e.g. le MTS décroît rapidement puis reste stable avant de croître de manière modérée). L'objectif est d'analyser non seulement le niveau du MTS mais également son rythme d'évolution, et donc les vitesses moyennes entre deux valeurs, puis de comparer ces deux représentations (niveau vs vitesse) dans le cadre de modèles d'extraction de motifs longitudinaux discriminants.

La vitesse moyenne du MTS entre deux temps consécutifs est définie comme la dérivée discrète d'ordre 1 :

$$vit_k = \frac{(val_k - val_{k-1})}{t_k - t_{k-1}}$$

Compte tenu de la régularité temporelle des valeurs, i.e. intervalles fixes de 6 mois, on utilise $t_k - t_{k-1} = 6$. Chaque patient est alors associé à un vecteur de vitesses moyennes $\mathcal{V}it = \{vit_1, vit_2, \ldots, vit_n)$

4.3 Discrétisation des valeurs et des vitesses

Afin de transformer les trajectoires numériques du MTS en représentations symboliques adaptées aux méthodes de fouille de motifs temporels symboliques telle que PrefixSpan, une discrétisation par quintiles a été appliquée.

L'ensemble des valeurs de MTS, tous patients et tous temps confondus, a été collectée afin d'estimer la distribution globale des valeurs et des vitesses moyennes. Les valeurs sont partitionnées en 5 catégories en fonction de leur position dans la distribution globale Q_{20}, Q_{40}, Q_{60}, Q_{80}, correspondant aux 20ième, 40ième, 60ième, et 80ième percentiles, respectivement. Chaque catégorie est alors associée à une étiquette décrivant :
(a) la direction et l'intensité de la vitesse moyenne du MTS : CR (Croissance Rapide), CM (Croissance Modérée), S (Stabilité), DM (Décroissance Modérée), DR (Décroissance Rapide).
(b) ou la magnitude des valeurs de MTS : VTE (Valeur Très Élevée), VE (Valeur Élevée), VM (Valeur Moyenne), VF (Valeur Faible), VTF (Valeur Très Faible).

En conséquence, pour chaque patient la représentation des vitesses moyennes successives devient une séquence $\mathcal{S}vit_p$=$(svit_1, \ldots, svit_{n_p})$ où $svit_i \in \{\text{DR}, \text{DN}, \text{S}, \text{CN}, \text{CR}\}$. De même, chaque patient sera associé à une séquence $\mathcal{S}val_p = (sval_1, \ldots, sval_{k_{n_p}})$ où $sval_i \in \{\text{VTF}, \text{VF}, \text{VM}, \text{VE}, \text{VTE}\}$.

4.4 Extraction de motifs séquentiels (PrefixSpan)

Nous appliquons une fouille de motifs sur les séquences symboliques, i.e $\mathcal{S}vit_p$ et $\mathcal{S}val_p$, afin d'identifier des sous-séquences récurrentes décrivant l'évolution temporelle du MTS. Nous nous appuyons sur l'algorithme *PrefixSpan*, qui extrait efficacement des motifs fréquents par croissance de préfixes via bases projetées, sans génération explicite de candidats [11]. Deux paramètres contrôlent l'extraction : un support minimal $s_{\min}$ (proportion minimale de patients contenant le motif) et une longueur minimale $L_{\min}$.

Nous considérons deux variantes de motifs : (i) **motifs non contigus** qui préservent l'ordre temporel avec éventuels écarts temporels entre les états, et (ii) **motifs contigus** (sous-chaînes) imposant une succession immédiate d'états. Chaque motif m induit ensuite une règle de décision $m \Rightarrow c$, où c est la classe, filtrée par des seuils de confiance $\gamma_{\min}$ et de lift $\lambda_{\min}$. Nous utilisons une implémentation open-source de PrefixSpan qui est disponible sur GitHub[2].

4.5 Extraction de motifs locaux numériques (*shapelets*)

Afin d'analyser finement la variation de MTS, nous appliquons l'extraction de *shapelets* aux deux représentations numériques suivantes : (1) la séquence des valeurs brutes du MTS harmonisée (i.e. $\mathcal{V}al = (val_1, val2, \ldots, val_{n_p})$) et la séquence des vitesses (i.e. $\mathcal{V}it = (vit_1, vit_2, \ldots, vit_n)$).

Un *shapelet* est une sous-séquence contiguë S extraite d'une série temporelle numérique, représentant un motif local caractéristique [9]. Contrairement aux motifs symboliques globaux, les shapelets permettent de capturer des séquences numériques locales discriminantes et de définir une distance minimum à ces séquences pour appartenir à la même classe. Soit S un shapelet de longueur $|S|$ et $\mathcal{X}_p$ une trajectoire patient telle que $\mathcal{X}_p \in \{\mathcal{V}al, \mathcal{V}it\}$ selon la représentation considérée. La présence du shapelet dans la trajectoire est évaluée par la distance euclidienne minimale z-normalisée :

$$d(S, \mathcal{X}_p) = \min_i \left\| z(S) - z\big(\mathcal{X}_p[i : i + |S| - 1]\big) \right\|_2 ,$$

où $z(\cdot)$ désigne la normalisation centrée-réduite et $\mathcal{X}_p[i : i + |S| - 1]$ une sous-séquence contiguë de même longueur que S. Chaque shapelet candidat induit une règle de la forme : $d(S, \mathcal{X}_p) \leq \tau \Rightarrow c$, où c est une classe cible et τ un seuil sélectionné par optimisation sous contraintes de support minimal, confiance minimale et lift minimal.

L'extraction est réalisée sur un ensemble borné de candidats (longueurs prédéfinies, sous-échantillonnage contrôlé). Nous utilisons les implémentations de référence des approches shapelets qui sont disponibles dans les bibliothèques open-source : `tslearn`[3] et `sktime`[4].

2. https://github.com/ekzhu/prefixspan
3. https://github.com/tslearn-team/tslearn
4. https://github.com/sktime

5 Découverte de règles FOL dans le graphe de connaissances KG_{LLC}

5.1 Graphe temporel de connaissances KG_{LLC}

Le recours à un graphe temporel de connaissances permet de structurer et d'intégrer des données hétérogènes (statiques, longitudinales et motifs) dans une représentation unifiée. Cette structuration est essentielle pour contextualiser les motifs discriminants en fonction des caractéristiques des patients et permettre l'apprentissage de règles explicables. Sans cette intégration, les motifs resteraient isolés et difficilement exploitables dans un cadre d'analyse interprétable. Un graphe de connaissances (GC) est constitué de trois composants : une ontologie, des vocabulaires contrôlés et les données couvertes par le graphe. La méthodologie de construction d'un GC comporte trois étapes [1] : - construction du corpus ; - construction de l'ontologie ; - instanciation du graphe pour répondre aux questions de compétence définies dans l'étape d'acquisition.

L'ontologie construite a pour objectif la validation de l'hypothèse du caractère précurseur du biomarqueur MTS pour prédire l'évolution de la LLC. Une version simplifiée de l'ontologie CLL Prognosis Ontology (CLLPO) [12] est utilisée car nous ne disposons pas de toutes les données relatives aux patients : les signes cliniques, les complications et les comorbidités. Les connaissances utilisées pour la construction de notre modèle sont relatives à la description du patient (âge, sexe, etc.), aux résultats des examens génétiques réalisés une seule fois sur les gènes à la date du diagnostic, aux prélèvements NFS et aux données MTS qui évoluent au cours du temps ainsi qu'aux traitements qui évoluent en fonction de l'état clinique du patient. Le concept de prélèvement généralise deux types de données (NFS et MTS) représentés par deux classes dans l'ontologie. Un prélèvement est caractérisé par une date exprimée en terme de mois et une valeur à cette date. Un traitement est défini par un nom et une date également exprimée en terme de mois. Ces concepts sont directement liés au patient par des relations permettant d'associer chaque patient aux données correspondantes. L'instanciation du graphe est réalisée en peuplant l'ontologie à partir des données collectées auprès des biologistes et nettoyées (fichier au format CSV) en utilisant le plugin Celfie de Protégé. La littérature et les discussions avec les chercheurs nous ont permis d'intégrer des données calculées utiles au moment de la génération de règles FOL : la classe d'âge et le doublement des lymphocytes qui traduisent une règle experte qui renseigne sur l'évolution de la maladie. Nous avons enrichi l'ontologie par la classe Motif qui se spécialise en trois sous-classes PrefixSpanNC, PrefixSpanC et shapelet afin de représenter pour chaque patient l'ensemble des motifs discriminants auxquels il peut être associé. Cette modélisation permet d'utiliser les motifs comme des descripteurs symboliques structurés dont la valeur et le type sont conservés pour être utilisés lors des étapes de découverte de règles FOL.

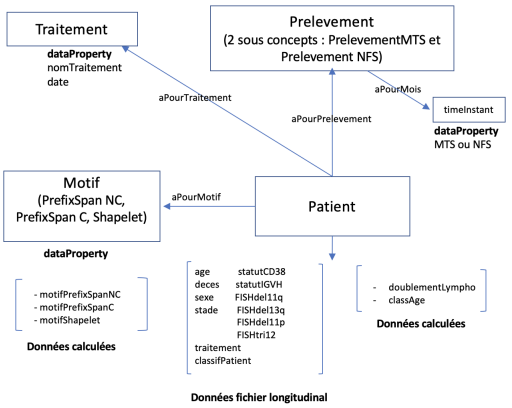


FIGURE 2 – Extrait de l'ontologie enrichie par les classes et les propriétés représentant les motifs

5.2 Découverte de règles fondée sur AMIE3

L'objectif est de découvrir un ensemble de règles fermées et connectées exprimées en logique du premier ordre (FOL) à partir de KG_{LLC} pouvant aider les biologistes à comprendre le rôle prédictif des mesures longitudinales du MTS sur la progression de la LLC. Ces règles peuvent également différencier ce rôle selon les caractéristiques statiques propres au patient (par ex. âge, sexe, tests génétiques) ou d'autres descripteurs dynamiques déjà connus. Les règles d'intérêt concluent sur la classe Progressive ou Indolente.

Règle de Horn connexe et fermée : Une règle de Horn $r : \mathcal{B} \Rightarrow H$ est une formule de logique du premier ordre où le corps de la règle $\mathcal{B}$ est une conjonction d'atomes $B_1, ..., B_n$ (c'est-à-dire des prédicats $B_i(X, Y)$ où X et Y sont des variables ou des constantes), et la conclusion, notée H, consiste en un seul atome représentant le fait qu'un patient B est progressif ou indolent. Une règle est dite fermée si chaque variable apparaît au moins deux fois dans la règle. Une règle est dite connectée si tous les atomes sont transitivement connectés, et deux atomes sont connectés s'ils partagent au moins une variable [7].

Exemple : La règle de Horn fermée et connexe suivante exprime le fait qu'un patient P de sexe féminin pour lequel le motif séquentiel discriminant "DR DR DR" identifié par m1 est couvert par sa séquence de valeur de vitesses moyennes de MTS, peut être classée comme progressive :

$Sexe(P, female), APourMotif(P, m1),$
$ValeurPrefixSpanC(m1, "DR\,DR\,DR")$
$\Rightarrow ClassifPatient(P, progressif)$

Nous distinguons parmi l'ensemble R des règles découvertes : les ensembles de règles R_{Static}, uniquement fondées sur les caractéristiques statiques du patient et R_{MTS} qui représentent l'ensemble des règles comportant au moins un motif en prémisse.

6 Expérimentation et évaluation

L'objectif expérimental est double : (i) vérifier que la prédiction de la forme clinique de la LLC (progressif, indolent) peut s'appuyer sur des *motifs décrivant la variation des*

vitesses du marqueur MTS plutôt que sur *la variation des valeurs normalisées* , et (ii) exploiter ces motifs discriminants pour produire des règles interprétables ciblant des sous classes de patients.

6.1 Caractéristiques du jeu de données

L'évaluation est menée sur une cohorte rétrospective longitudinale de **142** patients atteints de LLC, fournie par les biologistes de SIHMEL. Le fichier longitudinal brut comprend **1032** enregistrements cliniques horodatés. Après exclusion des patients sans étiquette d'évolution (statut manquant) ou disposant de moins de **3** mesures de MTS, le sous-ensemble conservé contient **102** patients, dont **57** progressifs (P) et **45** indolents (I). La cohorte présente ainsi un déséquilibre modéré des classes, ce qui motive l'usage de mesures macro-moyennées. Le jeu de données combine trois types d'informations complémentaires (voir table1) : (i) des **descripteurs statiques** (sexe, âge, statut IGHV et anomalies chromosomiques), (ii) des **biomarqueurs longitudinaux** (dont NFS et MTS), et (iii) des **variables dérivées ou ajoutées** (classe d'âge, doublement des lymphocytes, ...). Le nombre moyen d'observations MTS par patient est de **7,24** (min–max : 3–33), pour une durée médiane de suivi de **50** mois (Table 2).

TABLE 2 – Caractéristiques générales de la cohorte.

Caractéristique	**Valeur**
Nb de variables	21
Patients dont l'évolution est connue	102
Progressifs (P)	57
Indolents (I)	45
Observations MTS (tous patients, tous temps)	1023
Moyenne d'observations MTS par patient	7,24
Min–Max d'observations MTS par patient	3 – 33
Durée médiane de suivi (mois)	50,0
Nombre total de fenêtres de vitesse à 6 mois	1330

6.2 Évaluation du pouvoir discriminant des motifs MTS

6.2.1 Protocole d'évaluation

Nous adoptons un protocole **Leave-One-Out (LOO) sans fallback**. Pour chaque pli k : (1) le patient k est mis de côté pour le test ; (2) l'extraction de motifs et l'induction des règles sont réalisées *uniquement* sur les $N-1$ patients restants ; (3) le patient test n'est prédit *que si* au moins une règle s'applique. Lorsque plusieurs règles peuvent s'appliquer, la règle retenue est celle ayant la meilleure confiance (puis lift et support en cas d'égalité). Dans ce cadre, la méthode peut donc *s'abstenir* de prendre une décision. Les métriques utilisées sont les suivantes (N_{dec} étant le nombre de patients avec une décision, et N_{corr}, le nombre de correctes) :

$$\text{Cov} = \frac{N_{\text{dec}}}{N}, \quad \text{Acc}_{\text{cov}} = \frac{N_{\text{corr}}}{N_{\text{dec}}}, \quad \text{MacroF1} = \frac{\text{F1}_P + \text{F1}_I}{2}. \tag{1}$$

6.2.2 Résultats quantitatifs

Nous présentons les performances obtenues, en distinguant deux représentations : (i) les valeurs brutes du MTS et (ii) les vitesses du MTS calculées sur des fenêtres régulières de 6 mois (Tables 3 et 4).

Nous avons choisi les seuils permettant d'avoir le meilleur compromis couverture/accuracy : une règle est conservée si elle satisfait $s_{\min} = 0.18$ (support minimal), $L_{\min} = 2$ (longueur des motifs), $\gamma_{\min} = 0.70$ (confiance minimale) et $\lambda_{\min} = 1.30$ (lift minimal).

MTS valeurs brutes. Lorsque les modèles sont entraînés sur les valeurs normalisées du MTS (Table 3), les approches séquentielles présentent une couverture relativement faible : 0.35 pour PrefixSpan non contigu et 0.36 pour la version contiguë. Cependant, leur accuracy sur les patients couverts reste relativement élevée (0.66 et 0.73 respectivement), avec une Macro-F1 atteignant 0.84 pour la version contiguë. Les méthodes basées sur une représentation numérique (SAX-VSM et shapelets) offrent une couverture nettement supérieure (0.73 et 0.36), mais une accuracy plus modérée (0.61 et 0.66). WEASEL-lite montre une couverture limitée (0.15) mais une accuracy élevée sur les cas couverts (0.75), traduisant un comportement nettement plus sélectif.

Vitesses MTS à 6 mois. Lorsque l'on considère la dynamique du MTS via les vitesses moyennes semestrielles (Table 4), on observe une amélioration notable de la capacité discriminante des motifs non contigus : la couverture de PrefixSpan (NC) passe de 0.35 à 0.61 et son accuracy augmente de 0.66 à 0.71. Cette augmentation confirme que l'information prédictive est davantage portée par les variations de vitesse que par les valeurs normalisées. La version contiguë, bien que très précise (accuracy 0.78), reste fortement limitée en couverture (0.18). Les shapelets maintiennent une couverture élevée (0.81) avec une performance stable (accuracy 0.67). SAX-VSM conserve une couverture (0.18) mais avec une accuracy plus modérée (0.62). WEASEL-lite, quant à lui, présente une couverture très faible (0.26).

Comparaison du nombre de motifs : vitesses vs valeurs. L'analyse du nombre de motifs extraits met en évidence des différences importantes entre les deux représentations (cf. table 8). PrefixSpan non contigu extrait un nombre plus élevé de motifs discriminants pour les vitesses (15 contre 3 pour les valeurs). Même si ces motifs concluent très majoritairement sur la classe Indolent (12/15), la présence de 3 motifs évolutifs suggère une meilleure capacité à capter des dynamiques de progression que dans le cas des valeurs brutes.

Les shapelets, permettent de découvrir un nombre beaucoup plus important de motifs et en particulier quand les valeurs brutes sont considérées (64 contre 28 pour les vitesses).

Dans l'ensemble, la représentation basée sur les vitesses moyennes du MTS à 6 mois améliore particulièrement les performances des motifs séquentiels non contigus, confirmant que la dynamique d'évolution du MTS constitue un signal discriminant plus pertinent que les valeurs normalisées du MTS.

TABLE 1 – Variables disponibles (statiques, longitudinales, dérivées).

Catégorie	Variable	Description
Statiques	UPN	Numéro unique de patient (répétés sur les lignes longitudinales).
	Sexe	Variable démographique (F=1, M=2).
	Âge (au diagnostic)	Âge au moment du diagnostic.
	Âge (au premier MTS)	Âge lors de la première mesure MTS.
	Stade au diagnostic (Binet)	Stade clinique au diagnostic.
	Statut CD38	Marqueur moléculaire (POS=1, NEG=0).
	Statut IGHV	Statut mutationnel IGHV (UM=0, M=1)
	FISH del13q	Anomalie cytogénétique (binaire).
	FISH del11q	Anomalie cytogénétique (binaire).
	FISH tri12	Anomalie cytogénétique (binaire).
	FISH del17q	Anomalie cytogénétique (binaire).
	Statut décès	Indicateur de décès observé (Yes=1, No=0).
Longitudinales	MTS	MTS (%) .
	NFS (Lymphocytose)	Marqueur sanguin longitudinal leucémique.
Dérivées / annotées	Classification patient (finale)	Étiquette : *Progressif (P)* vs *Indolent (I)* (i.e variable cible).
	Traitement	Indicateur de présence (binaire)
	Doublementlympho	Doublement des lymphocytes (binaire)
	Début traitement (mois)	Temps (mois) avant initiation du traitement (si disponible).
	Survie (années)	Survie globale (diagnostic → décès ou fin d'étude).
	Survie >5 ans et >10 ans	Variables dérivées binaires.

TABLE 3 – LOO sans fallback – MTS valeurs à 6 mois.

Modèle	Cov.	Acc_{cov}	F1_{macro}
PrefixSpan (NC)	0.35	0.66	0.40
PrefixSpan (C)	0.36	0.73	0.42
shapelets	**0.73**	0.62	**0.66**
SAX-VSM	0.36	0.72	0.42
WEASEL-lite	0.38	**0.74**	0.30

TABLE 4 – LOO sans fallback – Vitesses MTS à 6 mois.

Modèle	Cov.	Acc_{cov}	F1_{macro}
PrefixSpan (NC)	**0.61**	0.71	**0.70**
PrefixSpan (C)	0.18	**0.78**	0.44
shapelets	0.60	0.69	0.69
SAX-VSM	0.18	**0.78**	0.44
WEASEL-lite	0.26	0.73	0.42

Exemples de règles extraites : Nous présentons quelques règles issues des trois modèles.
Les exemples présentés en table 5 illustrent une signature *Progressif* et une signature *Indolent* de forte confiance pour les motifs découverts par Prefix Span (cas des vitesses). La table 6 présente un motif contigu court mais de confiance élevée.

TABLE 5 – Exemples de motifs (PrefixSpan NC, vitesse)

Classe	Motif	L	Sup.	Conf.	Lift
Progressif	`DR S S S S`	5	19 (0.18)	0.84	1.51
Indolent	`D CN CR`	3	20 (0.19)	0.8	1.81

L'exemple de shapelet 141@0 :8 pour le patient 900141 présenté en tableau 7 signifie que : (1) 141 : le shapelet a été extrait à partir de la série d'un patient d'UPN 141, (2) @0 :8 : le shapelet S correspond à un segment de la série du patient 141 situé entre les positions 0 et 8. Ce shapelet a induit une règle de la forme $d(S, \mathcal{V}it_p) \leq 1.67 \Rightarrow Progressif$, permettant d'associer un patient à la classe progressif s'il possède une sous-séquence à distance $\leq$1.67.

TABLE 6 – Exemple de motif PrefixSpan contigu.

Classe	Motif (contigu)	L	Sup.	Conf.	Lift
Progressif	`DN CN`	2	19 (0.18)	0.78	1.78

TABLE 7 – Exemples de règles shapelets extraits. τ : seuil de distance (z-norm. euclidienne).

Classe	shapelet ID	L	τ	Conf.	Lift
Progressif	`141@0:8`	8	1.67	0.71	1.28
Indolent	`129@2:6`	4	1.31	0.81	1.83

6.2.3 Sensibilité aux seuils

Nous analysons l'impact des paramètres de filtrage (*support* $s_{\min}$, *confiance* $\gamma_{\min}$) sur le compromis accuracy–couverture pour les trois approches : PrefixSpan non contigu, PrefixSpan contigu et shapelets.

Sensibilité au seuil de confiance sur les vitesses MTS. Pour chaque valeur de $\gamma_{\min}$ (à partir de 0.70), nous présentons l'évolution de l'accuracy conditionnelle Acc_{cov} et de la couverture.
La Figure 3(a) montre que l'augmentation de $\gamma_{\min}$ améliore l'accuracy pour PrefixSpan non contigu et pour les shapelets (de 0.76 à 0.86 et de 0.70 à 0.79 respectivement, lorsque $\gamma_{\min}$ augmente de 0.70 à 0.80). À l'inverse, PrefixSpan contigu conserve une précision stable (0.78) entre $\gamma_{\min} = 0.70$ et $\gamma_{\min} = 0.75$, mais ne découvre plus de motifs dès que la

valeur atteint 0.80.

En parallèle, la Figure 4(a) montre une diminution nette de la couverture lorsque $\gamma_{\min}$ augmente. Pour PrefixSpan non contigu, la couverture diminue de 0.61 à 0.35 entre $\gamma_{\min} = 0.70$ et $\gamma_{\min} = 0.80$; pour les shapelets, elle chute de 0.60 à 0.23. PrefixSpan contigu est fortement limité en couverture dès $\gamma_{\min} = 0.70$ (Cov = 0.18) et s'abstient entièrement à partir de $\gamma_{\min} \geq 0.80$. Au-delà de $\gamma_{\min} \geq 0.85$, aucun des trois modèles ne prend de décision (Cov = 0). Ces résultats illustrent clairement le compromis inhérent au cadre sans fallback : des seuils plus stricts augmentent la précision des décisions retenues, mais réduisent mécaniquement la fréquence de décision.

Sensibilité au seuil de confiance sur les valeurs MTS. Pour chaque valeur de $\gamma_{\min}$, nous rapportons l'évolution de l'accuracy conditionnelle Acc_{cov} et de la couverture Cov.

Pour PrefixSpan non contigu, toutes les règles sélectionnées ont une confiance proche de 0.70. Dès que $\gamma_{\min}$ atteint 0.75, le modèle s'abstient entièrement (Cov = 0). L'accuracy conditionnelle est donc uniquement définie à $\gamma_{\min} = 0.70$ (0.66).

PrefixSpan contigu présente une meilleure stabilité : à $\gamma_{\min} = 0.75$, la couverture reste modérée (Cov = 0.54) avec une légère amélioration de la précision (Acc_{cov} = 0.75). Toutefois, au-delà de $\gamma_{\min} = 0.80$, aucune règle ne satisfait le seuil.

Les shapelets montrent un comportement plus progressif. La couverture diminue graduellement lorsque $\gamma_{\min}$ augmente (de 1.00 à 0.20), tandis que l'accuracy conditionnelle augmente fortement aux seuils élevés (Acc_{cov} = 0.88 pour $\gamma_{\min} \geq 0.85$), ce qui reflète la rareté des règles de très forte confiance.

La Figure 3(b) confirme cette dynamique : l'augmentation du seuil de confiance entraîne une réduction rapide de la couverture pour PrefixSpan, alors que les shapelets conservent une capacité décisionnelle jusqu'à des niveaux de confiance élevés.

Comparaison entre vitesses et valeurs du MTS. La comparaison des deux représentations met en évidence des différences importantes dans le comportement des modèles en fonction du seuil de confiance.

Pour PrefixSpan non contigu, les vitesses produisent une couverture plus élevée à $\gamma_{\min} = 0.70$ que les valeurs normalisées, dont la couverture est limitée à 0.35 dès ce seuil, avant de chuter rapidement à zéro. Elles présentent également une amélioration progressive de l'accuracy conditionnelle lorsque le seuil augmente.

Pour PrefixSpan contigu, un comportement similaire est observé : les vitesses maintiennent une capacité décisionnelle sur un intervalle de confiance plus large, tandis que les valeurs présentent une couverture plus faible et deviennent rapidement inopérantes lorsque $\gamma_{\min}$ augmente.

Concernant les shapelets, les valeurs normalisées présentent une couverture initialement plus élevée à $\gamma_{\min} = 0.70$ (Cov $\approx$ 0.74), mais celle-ci diminue fortement lorsque le seuil augmente (Cov $\approx$ 0.28 à $\gamma_{\min} = 0.80$), traduisant un filtrage plus sélectif des règles.

TABLE 8 – Nombre total de motifs extraits selon la classe prédite (Vitesses et valeurs).

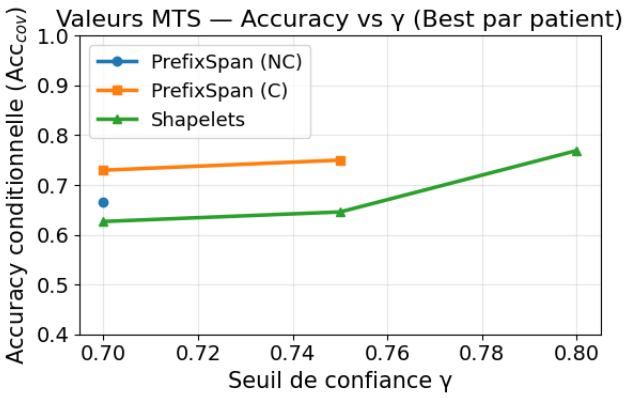
Modèle	Total	Progressif	Indolent
Vitesses			
PrefixSpan (NC)	15	3	12
PrefixSpan (C)	1	0	1
shapelets	28	6	22
Valeurs			
PrefixSpan (NC)	3	0	3
PrefixSpan (C)	2	2	0
shapelets	64	63	1

6.3 Génération des règles FOL et évaluation par les experts

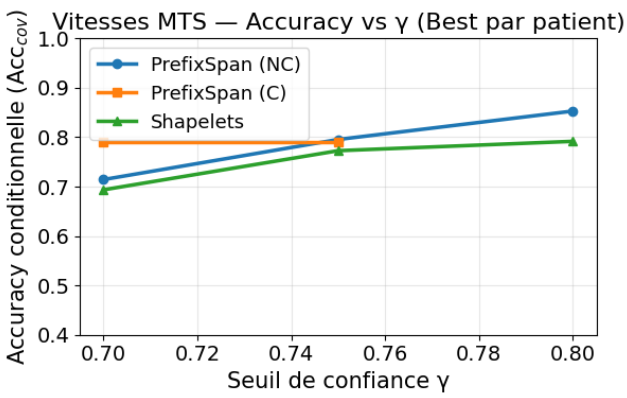
L'extraction de règles est réalisée à l'aide d'AMIE3 sur le graphe de connaissance peuplé et enrichi dans un premier temps par les motifs PrefixSpan sélectionnés. Plusieurs paramètres ont été fixés afin de contraindre l'espace de recherche et garantir la pertinence des règles générées. Nous restreignons les règles apprises à celles concluant sur la classe du patient (indolent/progressif) (`-htr`). Le support minimal (`-mins`) est fixé à 15, afin d'éliminer les règles reposant sur un nombre trop faible d'exemples positifs, tout en autorisant un support plus faible que les motifs seuls afin de les combiner avec d'autres informations. Le seuil minimal de confiance (`-minpca`) est fixé à 0.7 et le nombre d'atomes autorisés dans les prémisses est limité à 4, afin de contrôler la complexité structurelle des règles (`-maxad`). L'autorisation explicite de constantes dans les règles a été déclarée afin de capturer des associations impliquant des valeurs spécifiques (ex : sexe(?a,M)). Enfin, nous excluons du corps des règles certaines relations : (1) les prédicats non pertinents en prémisse tels que Décès, Age ou encore Stade et Traitement ; (2) les prémisses impliquant deux individus distincts liés par attributs ayant des variables identiques.

Résultats et première évaluation par les experts. AMIE3 a produit 63 règles comportant 3 ou 4 atomes. L'évaluation qualitative des règles a été réalisée par deux experts biologistes. Le protocole consiste à examiner la pertinence clinique des règles extraites et leur nouveauté. Plus précisément, Les experts ont évalué les 17 règles ne comportant pas de motif, en différenciant quatre catégories (cf. table 9) : connu (n=12), connu mais intéressant (n=1), intéressant (n=2), informe peu (n=1), pas forcément(n=1). Au cours de cette première phase d'évaluation, les experts ont également considéré le degré de confiance associé aux règles comportant des motifs (par exemple, R8 et R9) afin d'évaluer l'apport de cette information. Par la suite lorsque plusieurs experts seront impliqués, un accord inter-annotateurs pourra être mesuré afin d'évaluer la robustesse des annotations.

Des exemples de règles et d'annotation des règles sont représentés dans les tables 10 et 9. Par exemple, la règle R2 a été annotée comme "intéressante" dans la mesure où elle montre qu'en associant deux marqueurs cytogénétiques connus, la confiance peut être de 0.82. Parmi les règles comportant des

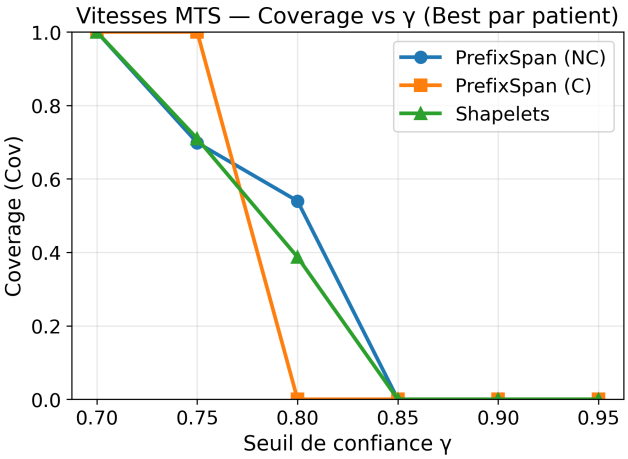


(a) Vitesses du MTS

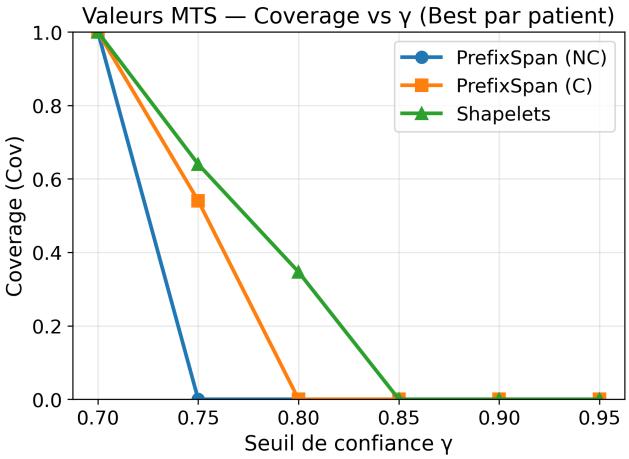


(b) Valeurs du MTS

FIGURE 3 – **Accuracy conditionnelle (Acc$_{cov}$) en fonction du seuil de confiance γ : comparaison entre une représentation dynamique (vitesses MTS) et une représentation statique (valeurs MTS).**



(a) Vitesses du MTS



(b) Valeurs du MTS

FIGURE 4 – **Couverture (Cov) en fonction du seuil de confiance γ : comparaison entre une représentation dynamique (vitesses MTS) et une représentation statique (valeurs MTS).**

motifs obtenus avec la technique PrefixSpan Non Contigu, 4 règles ont été annotées comme "extrêmement intéressante" car elles permettent d'identifier un motif associé aux patients progressifs. C'est le cas des règles R5, R6 et R7. Enfin, les règles telles que R8 montrent que le fait d'associer le marqueur FISHdel13q au motif augmente la confiance pour déterminer le caractère évolutif de la maladie par rapport au motif seul. L'ensemble des règles obtenues, en dehors de celles menant à des patients progressifs, associent des motifs à des patients indolents qui n'ont pour l'instant pas été évaluées par les experts (cas de la règle R9).

TABLE 9 – Distribution des annotations expertes pour les règles évaluées dans la table 10.

Règles	Catégorie	n	%	Confiance
–	Connue	12	70,6	–
R2	Connue mais intéressante	1	5,9	0,71
R1	Intéressante	1	5,9	0,82
R3	Informe peu	1	5,9	0,71
R4	Pas forcément	1	5,9	0,79
R5–R8	Extrêmement intéressante	4	0,06	[0,74, 0,84]

7 Discussion

Les résultats obtenus montrent que les motifs fondés sur les vitesses du MTS présentent un pouvoir discriminant supérieur à ceux fondés sur leurs valeurs normalisées. Cela suggère que la dynamique d'évolution du biomarqueur constitue une information plus pertinente pour caractériser la progression de la LLC. Cependant, plusieurs limitations doivent être soulignées. Tout d'abord, la taille de la cohorte reste limitée, ce qui peut impacter la généralisation des résultats. Ensuite, l'hypothèse d'interpolation linéaire constitue une approximation simplificatrice. Enfin, le pipeline proposé repose sur plusieurs paramètres pouvant influencer les motifs extraits.

TABLE 10 – Exemples représentatifs de règles annotées par les experts (sans et avec motifs).

ID	Règle
R1	$statutCD38(?a, CD38\_NEG), statutIGVH(?a, UM) \Rightarrow classifPatient(?a, progressif)$
R2	$FISHtri12(?a, tri12\_NEG), sexe(?a, F), statutIGVH(?a, M) \Rightarrow classifPatient(?a, indolent)$
R3	$FISHdel13q(?a, DEL13q\_NEG), sexe(?a, M), statutCD38(?a, CD38\_NEG) \Rightarrow classifPatient(?a, progressif)$
R4	$statutIGVH(?a, UM) \Rightarrow classifPatient(?a, progressif)$
R5	$aPourMotif(?a, ?f), valueMotifPS\_NC(?f, DR\,S\,S\,S\,S) \Rightarrow classifPatient(?a, progressif)$
R6	$aPourMotif(?a, ?f), valueMotifPS\_NC(?f, S\,S\,S\,S\,S) \Rightarrow classifPatient(?a, progressif)$
R7	$aPourMotif(?a, ?f), valueMotifPS\_NC(?f, DR\,S\,S\,S) \Rightarrow classifPatient(?a, progressif)$
R8	$FISHdel13q(?a, DEL13q\_NEG), aPourMotif(?a, ?h), valueMotifPS\_NC(?h, DR\,S\,S\,S) \Rightarrow classifPatient(?a, progressif)$
R9	$aPourMotif(?a, ?h), lymphoDouble(?a, DL0), valueMotifPS\_NC(?h, DR\,DR\,CR\,D) \Rightarrow classifPatient(?a, indolent)$

8 Conclusion

L'approche proposée permet d'explorer différents types de motifs pour évaluer la valeur prédictive du MTS des formes indolentes et progressives de la LLC. Les résultats montrent que des motifs discriminants peuvent être découverts aussi bien à partir de séquences harmonisées des valeurs du MTS que des séquences de vitesses moyennes. De plus, le caractère prédictif du MTS peut être évalué seul ou en conjonction avec d'autres données du patient. Bien sûr, d'autres expérimentations doivent être menées afin d'enrichir l'ontologie par les Shapelets pour expliciter des règles concluant sur la classe Progressif plus nombreuses. Les évaluations et les interfaces permettant d'interpréter les motifs discrétisés ou numériques doivent être développées. Enfin, l'instant auquel un motif discriminant est couvert devrait être explicité.

Références

[1] M. Cochez C. d'Amato G. de Melo C. Gutierrez S. Kirrane G. Labra J.E.L. Gayo R. Navigli-S. Neumaier A.C.N. Ngomo A. Polleres S.M. Rashid Sabbir M. A. Rula L. Schmelzeisen J. Sequeda S. Staab A. Zimmermann A. Hogan, E. Blomqvist. Knowledge graphs. *ACM Computing Surveys*, 54(4), 2021.

[2] Ferner R. E. Aronson, J. K. Biomarkers—a general review. *Current Protocols in Pharmacology*, 76 :9.23.1–9.23.17, 2017.

[3] N. Chevallier T. Beitar V. Eclache M. Quettier M. Boubaya R. Letestu V. Levy F. Ajchenbaum-Cymbalista N. Varin-Blank C. Le Roy, P.A Deglesne. The degree of bcr and nfat activation predicts clinical outcomes in chronic lymphocytic leukemia. *Blood*, 120(2) :356–365, 2012.

[4] Josif Grabocka, Nicolas Schilling, Martin Wistuba, and Lars Schmidt-Thieme. Learning time-series shapelets. In *Proceedings of the 20th ACM Conference on Knowledge Discovery and Data Mining*, pages 392–401, 2014.

[5] Seyed Mehran Kazemi, Rishab Goel, Kshitij Jain, Ivan Kobyzev, Akshay Sethi, Peter Forsyth, and Pascal Poupart. Representation learning for dynamic graphs : A survey. *Journal of Machine Learning Research*, 21(70) :1–73, 2020.

[6] Armita Khajeh Nassiri, Nathalie Pernelle, and Fatiha Saïs. Regnum : Generating logical rules with numerical predicates in knowledge graphs. In *Proceedings of the Extended Semantic Web Conference*, 2023.

[7] Julien Lajus, Luis Galárraga, and Fabian M. Suchanek. Fast and exact rule mining with amie 3. In *Proceedings of the Extended Semantic Web Conference*, 2020.

[8] Jessica Lin, Eamonn Keogh, Li Wei, and Stefano Lonardi. Experiencing sax : A novel symbolic representation of time series. *Data Mining and Knowledge Discovery*, 15(2) :107–144, 2007.

[9] Jason Lines, Luke M. Davis, Jon Hills, and Anthony Bagnall. A shapelet transform for time series classification. In *SIGKDD*, pages 289–297, 2012.

[10] Christian Meilicke, Melisachew Wudage Chekol, Daniel Ruffinelli, and Heiner Stuckenschmidt. Anytime bottom-up rule learning for knowledge graph completion. In *IJCAI*, pages 3137–3143, 2019.

[11] Jian Pei, Jiawei Han, Behzad Mortazavi-Asl, Jianyong Wang, Helen Pinto, Qiming Chen, Umeshwar Dayal, and Mei-Chun Hsu. Mining sequential patterns by pattern-growth : The prefixspan approach. *IEEE Transactions on Knowledge and Data Engineering*, 16(11) :1424–1440, 2004.

[12] C Piriou, S Despres, Nobecourt J, C Le Roy, C Irles, F Baran-Marszak, and Lévy V. Knowledge graph and ontology for representing cll data. In *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2023.

[13] Marco Postiglione, Daniel Bean, Zeljko Kraljevic, Richard J. B. Dobson, and Vincenzo Moscato. Predicting future disorders via temporal knowledge graphs and medical ontologies. *IEEE Journal of Biomedical and Health Informatics*, 28(7) :4238–4248, 2024.

[14] Patrick Schäfer and Ulf Leser. Fast and accurate time series classification with weasel. In *Proceedings of the 2017 ACM CIKM*, pages 637–646, 2017.