

Identification interactive de sous-groupes pour l'aide au diagnostic de la sclérose latérale amyotrophique

Mots clefs : fouille de données, aide au diagnostic

Sous-axes de la fédération ICVL concernés : A1 (Knowledge extraction and pattern recognition)

RTR DIAMS : Axe 3 : Apprentissage, Optimisation et Aide à la décision / GT: Human In the Loop for Data mining and Machine Learning

Equipes partenaires : LIFAT, LIFO, iBrain (UMR 1253)

Budget demandé : financement de cinq à six mois de stages de master 2

Contexte

La sclérose latérale amyotrophique (SLA) est une affection neurodégénérative qui se caractérise par une paralysie progressive des muscles conduisant à une perte de l'élocution, de la déglutition, de la mobilité des membres et à une paralysie des muscles respiratoires. Dans la majorité des cas, ces paralysies conduisent à un décès après une évolution médiane de 36 mois, avec des cas extrêmes de quelques mois à plusieurs années. La physiopathologie de la SLA reste complexe avec différents mécanismes tels que l'excitotoxicité médiée par le glutamate ou l'agrégation anormale des protéines, mais elle n'est pas suffisamment bien documentée pour 1) permettre un diagnostic précoce, 2) suggérer un pronostic ou encore 3) proposer une thérapeutique ciblée efficace. Finalement, un seul traitement, le riluzole, a montré une efficacité objectivée par une prolongation de la survie de 2 à 3 mois en moyenne.

Ainsi, la prise en charge de la SLA est délicate dès le diagnostic retardé d'environ 1 an par rapport aux premiers symptômes. La recherche de biomarqueurs dans la SLA a pris une place capitale et a incité à l'utilisation d'outils hauts débits d'exploration biologiques. Néanmoins, malgré la quantité importante de données cumulées, il n'existe à ce jour aucun biomarqueur diagnostique ou capable de discriminer les profils cliniques.

Dans le cadre d'un stage master financé par l'ICVL et réalisé au printemps 2019, une base de données PostgreSQL intégrant l'ensemble des données cliniques sur 1257 patients atteints de la SLA et suivis au sein du CHU de Tours a été mise en place. Pour chaque patient, cette base contient en particulier l'évolution de leur poids, scores ALS, CVL, CVF,... De plus, cette base intègre de nombreuses données biologiques (issues d'analyses de prises de sang au cours du suivi des patients). L'intégration de données étant terminée, le projet proposé portera alors sur l'analyse et la fouille des données collectées pour l'aide au diagnostic et suivi des patients. Dans le cadre d'un autre stage de master financé par l'ICVL en 2020, une représentation de données longitudinales ainsi qu'une mesure de distances ont été proposées. Un premier travail pour identifier des sous-groupes pertinents a été réalisé. Les résultats préliminaires nous ont permis de confirmer le rôle de certains acteurs biologiques dans la maladie en cohérence avec ce qui a été précédemment décrit et identifié par des approches d'analyses statistiques standards. De plus, les outils d'analyse de

données longitudinales tels que récemment développés nous permettent d'envisager la poursuite de ces stratégies peu explorées à ce jour.

Description du projet

Depuis quelques années, Hélène Blasco et Philippe Corcia (iBrain, UMR 123) ont conduit différents travaux et expérimentations pour évaluer l'apport des techniques de fouille de données à l'identification de biomarqueurs, à la fois pour l'aide au diagnostic de la SLA [1,2,3] et pour l'aide à la prédiction de ces évolutions [4,5]. De nombreuses expérimentations ont également été menées [9] dans le cadre d'un consortium international et de l'analyse des données de la plateforme PROACT (Pooled Resource Open-Access ALS Clinical Trials Database).

Il reste néanmoins de nombreux challenges concernant l'aide à l'identification de sous-groupes de patients présentant des évolutions cliniques similaires, mais aussi pour aider à l'identification d'ensembles de biomarqueurs pour l'aide à la prédiction. L'objectif du stage est d'étudier un mécanisme pour intégrer l'expert qui fournit les connaissances dans la boucle d'apprentissage et pour intégrer des contraintes ou préférences de l'expert. L'interactivité avec l'expert pourra se faire par le retour de l'expert sur le résultat, par exemple le veto si le résultat représente l'inverse de ce qui est une tendance dans la littérature ou dans les données avec les types d'analyse existants. La méthode choisie devra être testée et faire l'objet d'expérimentations précises en termes de validation des résultats obtenus. Dans ce cadre, ce stage s'appuiera sur les travaux récents des équipes « Bases de données et traitement des langues naturelles » du LIFAT [6] et « Contraintes et apprentissage » du LIFO [7,8], et sur l'expertise des médecins de l'équipe iBrain.

Références

- [1] Blasco H., Nadal-Desbarats L., Pradat P.F., et al. Untargeted 1 H-NMR metabolomics in CSF: toward a diagnostic biomarker for motor neuron disease. *Neurology*; 82: 1167–1174 (2014).
- [2] Blasco H. et al. Comparative analysis of targeted metabolomics: Dominance-based rough set approach versus orthogonal partial least square-discriminant analysis. *Journal of Biomedical Informatics*, 53, pp. 291-299 (2015).
- [3] Blasco, H. , Nadal-Desbarats, L. , Pradat, P. , Gordon, P. H., Madji Hounoum, B. , Patin, F. , Veyrat-Durebex, C. , Mavel, S. , Beltran, S. , Emond, P. , Andres, C. R. and Corcia, P. Biomarkers in amyotrophic lateral sclerosis: combining metabolomic and clinical parameters to define disease progression. *European Journal of Neurology*, 23: pp. 346-353 (2016).
- [4] Blasco H., F Patin F., S Molinier S., P Vourc'h P., O Le Tilly O., Bakkouche S., C R Andres C. R., Meininger V., Couratier P. & Corcia P., A decrease in blood cholesterol after gastrostomy could impact survival in ALS. *European Journal of Clinical Nutrition* volume 71, pp. 1133–1135 (2017).
- [5] Patin, F. , Corcia, P. , Madji Hounoum, B. , Veyrat-Durebex, C. , Respaud, E. , Piver, E. , Benz-de Bretagne, I. , Vourc'h, P. , Andres, C. R. and Blasco, H., Biological follow-up in amyotrophic lateral sclerosis: decrease in creatinine levels and increase in ferritin levels predict poor prognosis. *European Journal of Neurology*, 22: pp. 1385-1390 (2015).
- [6] A. El Moussawi, A. Cheriati, A. Giacometti, N. Labroche, A. Soulet. Clustering with Quantitative User Preferences on Attributes. *ICTAI 2016*: 383-387.

[7] T. A. Lampert, T.-B.-H. Dao, B. Lafabregue, N. Serrette, G. Forestier, B. Crémilleux, C. Vrain, P. Gançarski, Constrained distance based clustering for time-series: a comparative and experimental study. *Data Mining Knowledge Discovery*, 32(6): pp. 1663-1707 (2018).

[8] T.-B.-H. Dao, K.-C. Duong, C. Vrain, Constrained clustering by constraint programming. *Artificial Intelligence*, 244: pp. 70-94 (2017).

[9] R. Kueffner et al. "Stratification of amyotrophic lateral sclerosis patients: a crowdsourcing approach." *Scientific reports* 9.1 (2019): 690.

Personnes à contacter/Encadrants

- Arnaud Giacometti et Arnaud Soulet (LIFAT)
- Thi-Bich-Hanh Dao et Christel Vrain (LIFO)
- Hélène Blasco et Philippe Corcia (iBrain)