Approche préventive de la qualité des données d'importation dans le contexte de la protéomique clinique

Pierre Naubourg, Marinette Savonnet, Éric Leclercq et Kokou Yétongnon

Université de Bourgogne, Laboratoire LE2I - UMR5158
9 Avenue Alain Savary
21000 DIJON, FRANCE
{pierre.naubourg, marinette.savonnet, eric.leclercq, kokou}@u-bourgogne.fr
http://le2i.cnrs.fr

Résumé. Dans le domaine biomédical, la protéomique est confrontée à des sources de données de plus en plus nombreuses et à des volumes de données très importants du fait de la multiplication des technologies dites à haut débit. L'hétérogénéité de la provenance des données implique de fait une hétérogénéité dans la représentation et le contenu de ces données. Les données peuvent aussi se révéler incorrectes ce qui engendre des erreurs sur les conclusions des expériences protéomiques. Notre approche a pour objectif de garantir la qualité initiale des données lors de leur importation dans un système d'information dédié à la protéomique. Elle est basée sur le couplage entre des modèles représentant les sources et le système protéomique, et des ontologies utilisées comme médiatrices entre les modèles. Les différents contrôles que nous proposons de mettre en place garantissent la validité des domaines de valeurs, la sémantique et la cohérence des données lors de l'importation.

1 Introduction

Notre contexte de travail est le domaine biomédical et plus précisément la protéomique clinique. La particularité de la protéomique clinique est la recherche de caractéristiques protéiniques d'échantillons issus de groupes de patients participant à une étude. Parmi ces caractéristiques, nous pouvons donner comme exemple la découverte de biomarqueurs permettant d'identifier une pathologie et ainsi de la classifier, d'effectuer un diagnostic précoce, d'étudier la réponse du patient au traitement, etc. Le travail des plateformes protéomiques est centré sur la réalisation d'études mettant en jeu un grand nombre d'échantillons dont on essaie d'extraire des caractéristiques via des expérimentations. Outre les données nécessaires à l'analyse des échantillons sur les spectromètres de masse par exemple, la réalisation d'études statistiques en aval de ces expériences nécessite l'utilisation de données cliniques. Les données cliniques englobent des données aussi larges que les caractéristiques du patient, la description des pathologies diagnostiquées, les caractéristiques des échantillons prélevés, les conditions de transport et de stockage.