

Extraction d'informations sur les workflows scientifiques à partir de la littérature

Clémence Sebe*, Aurélie Névéal*, Sarah Cohen-Boulakia*, Alban Gaignard**

*Université Paris-Saclay, CNRS, LISN
prenom.nom@lisn.fr

**Université de Nantes, CNRS, INSERM Institut du Thorax
alban.gaignard@univ-nantes.fr

Résumé. Les workflows scientifiques offrent aux bioinformaticiens un cadre pour représenter, échanger et assurer la reproductibilité de leurs pipelines d'analyses. Ils sont décrits dans la littérature (texte) et/ou stockés dans des dépôts de workflows (code). Un enjeu majeur pour tendre vers une meilleure réutilisation des workflows par des tiers est de reconstruire le lien entre la documentation (texte) et l'implémentation (code) du workflow. A partir du texte intégral d'articles décrivant des workflows en anglais, nous proposons une méthode de modélisation et d'extraction d'informations des composants des workflows. Nous présentons un corpus de 24 articles annotés à l'aide d'un schéma comportant 16 entités et 10 relations. Nous utilisons ce corpus pour entraîner et évaluer des modèles statistiques d'extraction d'information sur les workflows. Nous montrons la faisabilité de la tâche comme première étape vers l'intégration d'information concernant les workflows issus de la littérature et des dépôts de workflows.

1 Introduction

Dans les sciences fortement génératrices de données, telles que la *bioinformatique*, les résultats scientifiques sont produits à l'aide de chaînes de traitement complexes pouvant prendre en entrée de très grandes quantités de données expérimentales. Ces chaînes peuvent être constituées de nombreuses étapes, faire appel à des outils bioinformatiques, et demander des temps de calcul considérables. Elles peuvent être implémentées à l'aide de scripts (Bash, Python...) qui pilotent l'exécution des outils, et constituent le liant entre les données, les traitements, les outils et l'environnement d'exécution. Cependant, le développement et l'utilisation de tels scripts engendrent de multiples difficultés d'implémentation dans la conception et l'exécution de la chaîne de traitement ainsi que des difficultés de maintenance et réutilisation par des tiers.

En réponse à ces problèmes, des efforts considérables ont été déployés ces vingt dernières années pour proposer aux bioinformaticiens des approches pour les guider vers une meilleure automatisation de leurs chaînes de traitement : les *systèmes de workflows scientifiques* (Cohen-Boulakia et al. (2017)). Deux systèmes sont aujourd'hui particulièrement utilisés en bioinformatique : Nextflow (Di Tommaso et al. (2017)) et Snakemake (Köster et Rahmann (2012)).

On assiste actuellement à deux phénomènes : (1) augmentation des articles scientifiques décrivant des workflows bioinformatiques qui sont sous une forme "descriptive", (étapes dé-