SUJET DE THÈSE

Analyse de données multi-modales pour les Pathologies complexes par la conception et l'implémentation de Protocoles Reproductibles et Réutilisables

Encadrement
Sarah Cohen-Boulakia, Université Paris-Saclay (directrice de thèse);
Alain Denise, Université Paris-Saclay
Alban Gaignard, Université de Nantes (co-encadrant principal)

Profil étudiant recherché : Master M2 en Informatique ou Bioinformatique. Bonnes connaissances en bases de données (si possible en intégration de données), représentation des connaissances (RDF), algorithmique des graphes. Programmation Python. Très bonnes capacités à communiquer notamment dans un milieu interdisciplinaires. Capacité à échanger en anglais est un plus. Connaissances de systèmes de workflows scientifiques (NextFlow, SnakeMake, Galaxy...) est un plus.

1 - Résumé

L'étude de pathologies comme les anévrismes intracrâniens nécessite l'utilisation d'une grande variété de données et la conception de protocoles d'analyse complexes. La diversité de leurs implémentations rend leur maintenance et partage difficile et limite la confiance des biologistes dans les données produites. Reproduire et réutiliser les protocoles est pourtant crucial pour comparer systématiquement les résultats biologiques, adapter des protocoles à de nouvelles problématiques et répondre aux exigences des *plans de gestion de données*. L'objectif de cette thèse est de fournir (i) une large bibliothèque de protocoles organisés, (ii) un module de conception et d'exécution de protocoles reproductibles, réutilisables et citables (conception d'algorithmes d'indexation et de recherche efficace de motifs dans les graphes formés par les workflows implémentant les protocoles), (iii) une évaluation de l'approche et (iv) un ensemble de critères FAIR pour les protocoles.

Ce sujet est financé par le CNRS (projet R2P2, appel 80 prime) où le/la doctorant.e collaborera avec des chercheurs du Laboratoire de Recherche en Informatique (LRI, Saclay) et de l'Institut du Thorax (ITX, Nantes).

Mots-clés

Intégration de données biologiques, Réutilisation et échange de protocoles, workflows scientifiques et protocoles FAIR, Analyse de données multi-échelles.

Contexte et état de l'art

L'anévrisme intracrânien est une anomalie vasculaire cérébrale affectant 3,2% de la population Française. Alors que sa rupture peut conduire au décès ou à un handicap sévère, il n'y a aucun outil diagnostic. L'étude de ces pathologies nécessitent i) l'utilisation d'une grande variété de jeux de données acquises à différentes échelles (génome, tissus vasculaires, organe vasculaire cérébral, population) dans le cadre de collaborations multidisciplinaires et multi-site et ii) la conception de protocoles d'analyse complexes et variés. Il est crucial de pouvoir reproduire ces analyses avec un fort niveau de confiance sur des jeux de données. Cependant, **le partage de données de santé est souvent freiné par les impératifs de protection des données personnelles et se heurte à des contraintes techniques (sécurité, volume).** Ces contraintes peuvent cependant être limitées lorsque les protocoles sont suffisamment réutilisables pour reproduire des analyses *in situ*. Aussi, lorsqu'ils sont conçus pour être réutilisables, les implémentations de protocoles (ou workflows) fournissent la provenance des données analysées, et augmentent la confiance des scientifiques dans les résultats produits.

La reproductibilité et réutilisation de protocoles doit faire face à de nombreux défis. C'est lorsqu'un protocole est reproductible qu'il peut être échangé pour être réutilisé en totalité ou partie, ou adapté pour répondre à de nouvelles questions biologiques. La crise de la reproductibilité qui a éclaté il y a 15 ans [SPZA03, AQM+11] a mis en évidence l'incapacité à reproduire des résultats obtenus par des méthodes bioinformatiques pour des raisons très diverses (manque de documentation sur les outils utilisés, non disponibilité des bibliothèques...). Une série de bonnes pratiques a vu le jour, combinées au développement de systèmes capturant la provenance des outils, jeux de données et informations relatives à l'environnement [DCE+07, Boe15, GNT10, BCC+13].
Néanmoins, les protocoles sont conçus et implémentés sans cadre adapté. Les systèmes de workflows offrent des interfaces de développement mais aucun ne permet de garder la trace des workflows réutilisés lors de la construction d'un nouveau workflow. **Il en résulte un nombre croissant de workflows dérivés de workflows pré-existants. Il est donc difficile d'identifier l'origine d'un protocole et de son implémentation et de maintenir les nombreuses implémentations de ces protocoles de façon cohérente et efficace.**
Alors que de nombreux travaux se sont attaqués à la production de données FAIR (Findable Accessible Interoperable Reusable) [WDA+16, MNV+17, HKP+18], le concept central de protocoles FAIR n'a été considéré que très récemment [GSS+20, Fai20]. Les principes FAIR [WDA+16] doivent être étendus pour prendre en compte notamment le caractère modulaire des protocoles et de leurs implémentations.

Verrous scientifiques, Objectifs et Méthodologie.
L'objectif de cette thèse est double (i) concevoir un cadre pour la conception et l'implémentation de protocoles d'analyse de données reproductibles et réutilisables pour l'étude des anévrismes intracrâniens et (ii) démontrer l'intérêt de cette approche en ré-utilisant et adaptant les protocoles obtenus en (i) sur des données générées dans de nouveaux projets. Sur le plan informatique, cette thèse apportera des solutions à la définition de protocoles FAIR avec des contributions relatives à la conception i) d'algorithmes d'indexation et de recherche efficace de motifs dans les graphes formés par les workflows et ii) la conception et l'implémentation d'outils d'aide à la réutilisation (et à la citation) de workflows. Sur le plan applicatif, cette thèse fournira des solutions concrètes pour documenter automatiquement les données produites par les protocoles annotés, tel qu'attendu dans un *Data Management Plan*. Elle débouchera sur un cadre d'échange de protocoles compréhensibles par les pairs qui démontrera sa capacité à réutiliser et adapter facilement des protocoles complexes développés dans un projet sur les données d'un nouveau projet.

Tâche 1 : Base de protocoles réutilisables pour l'étude des anévrismes intracrâniens. L'objectif est d'effectuer un recensement des protocoles et implémentations pour effectuer des analyses de données pour l'étude des anévrismes intracrâniens. T1.1 recense les protocoles déjà en place à l'ITX et collaborateurs avec un focus sur les protocoles exploitant des données biologiques. D'autres sources d'informations seront exploitées : (i) catalogues d'outils et de workflows, (ii) catalogues de code open-source de logiciels et workflows (e.g., GitHub), (iii) articles de la littérature (sections méthodologiques). T1.2 vise à semi-automatiser l'exploitation des sources de T1.1. T1.3 représente et organise les protocoles en vue de leur réutilisation. Cette tâche s'appuie sur des formalismes compatibles avec les principes FAIR et exploitables en algorithmique des graphes.

Tâche 2 : Module de réutilisation de protocoles reproductibles. L'objectif est de développer un module de réutilisation de protocoles reproductibles reposant sur un algorithme de traçage des (sous-)protocoles (ré)utilisés lors de la conception d'un nouveau protocole. Les défis sont nombreux : les protocoles et leurs implémentations forment des graphes complexes annotés par de nombreux termes issus d'ontologies. On définira un mécanisme (i) d'indexation des briques de base présentes dans les protocoles (vues comme des motifs dans des graphes) pour identifier ces briques de façon compacte et informative et (ii) de reconstruction de l'histoire d'un protocole, dont le problème est directement lié à celui de comparaison de ces protocoles

(sachant que le problème d'isomorphisme de sous-graphes est *difficile algorithmiquement*). T2.2 implémentera concrètement la méthode T2.1 qui s'applique ici aux systèmes largement supportés par la communauté (Galaxy, SnakeMake ou NextFlow) et en se fondant sur des standards de spécification de workflows (e.g. CWL).

Tâche 3 : Réutilisation de workflows pour l'analyse de nouvelles données. T3 a pour but de conduire une évaluation des résultats obtenus sur les protocoles de T1 par les algorithmes conçus et implémentés en T2. T3 est menée en étroite collaboration avec les chercheurs cliniciens et biologistes ayant fourni des protocoles d'analyse. T3.1 considère des protocoles conçus dans leurs projets passés et re-exécutés sur les données de nouveaux projets de l'ITX. Une adaptation des protocoles à ces nouvelles données sera aussi évaluée dans le cas où les caractéristiques des nouveaux jeux de données nécessitent de repenser certaines étapes. Les nouvelles données obtenues seront interprétées en lien étroit avec les biologistes de l'ITX. T3.2 propose des métriques pour évaluer la capacité d'un protocole ou workflow à être réutilisé ou reproductible, contribuant ainsi aux travaux de la communauté pour définir des protocoles FAIR.

Résultats attendus. (i) Bibliothèque de workflows annotés exécutables d'analyse de données dans le contexte des anévrismes intracrâniens, (ii) Module d'indexation et de citation de workflows, (iii) Evaluation de la robustesse des résultats biologiques obtenus par protocoles réutilisés, (iv) Critères FAIR dédiés aux protocoles.

Bibliographie

[AQM+11] A. Alsheikh-Ali, and W. Qureshi and Mouaz A. et al. Public availability of published research data in high-impact journals, PloS one, 6(9):e24357, 2011, Public Library of Science

[BCC+13] K. Belhajjame, J. Cheney, D. Corsar *et al.*, PROV-O: The PROV Ontology, W3C recommendation (2013)

[BCB+17] R. Bourcier, S. Chatel,..., **A. Gaignard** et al. on behalf of the ICAN Investigators, Understanding the Pathophysiology of Intracranial Aneurysm: The ICAN Project, Neurosurgery, 80(4):621–626, 2017

2018, Pages 133-141, ISSN 0002-9297, https://doi.org/10.1016/j.ajhg.2017.12.006

[Boe15] C. Boettiger, An introduction to Docker for reproducible research, ACM SIGOPS Operating Systems Review, 49(1):71--79,2015

[PAC+17] C Pradal, S Artzet, J Chopard… **S. Cohen-Boulakia**, InfraPhenoGrid: a scientific workflow infrastructure for plant phenomics on the grid, Future Generation Computer Systems 67, 341-353, 2017

[CBG+17] S. Cohen-Boulakia,..., **A. Gaignard** et al. Scientific workflows for computational reproducibility in the life sciences: Status, challenges and opportunities. Future Generation Computer Systems, 75, 284-298, 2017.

[DCE+07] S. Davidson, **S. Cohen-Boulakia,** A. Eyal et al. Provenance in Scientific Workflow Systems. IEEE Data Eng. Bull.,30(4):44--50, 2007

[DCF+17] P. Di Tommaso, M Chatzou, EW Floden et al. Nextflow enables reproducible computational workflows, Nature biotechnology 35(4):316, 2017

[Fai20] Fair workflows project (starting) https://fair-workflows.github.io/project.html

[GNT10] J. Goecks, A. Nekrutenko, J. Taylor, Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences, Genome Biology, 11(8),R86,2010

[GBSm+17] A. Gaignard, K. Belhajjame, H. Skaf-Molli. SHARP: Harmonizing and Bridging Cross-Workflow Provenance. ESWC 2017 Satellite Events Revised Selected Papers, 2017. <hal-01768394>

[GSmB+19] A. Gaignard, H. Skaf-Molli, K. Belhajjame, Findable and Reusable Workflow DataProducts: A Genomic Workflow Case Study, Semantic Web Journal, Special Issue on Semantic e-Science, 2019 (accepted)

[GSS+20] C. Goble, **S. Cohen-Boulakia**, et al. FAIR computational workflows. Data Intelligence, 108-121, 2020

[HKP+18] P. Holub, F. Kohlmayer, F. Prasser et al. Enhancing reuse of data and biological material in medical research: From FAIR to FAIR-health. Biopreservation and biobanking, 16(2):97-105, 2018.

[KR12] J. Köster and S.Rahmann. Snakemake - A scalable bioinformatics workflow engine. Bioinformatics 2012.

[LCL+19] F. Lemoine, D. Correia, V. Lefort… **S. Cohen-Boulakia**, O. Gascuel, NGPhylogeny.fr: new generation phylogenetic services for non-specialists, Nucleic Acids Research, 47(W1):W260–W265

[MNV+17] B. Mons, C Neylon, J Velterop et al. Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud, Information Services & Use 37 (1):49-56, 2017.

[NAB+19] A. Nouri, F. Autrusseau, R. Bourcier, …, **A. Gaignard,** *et al.* 3D bifurcations characterization for intra-cranial aneurysms prediction, Proc. SPIE 10949, Medical Imaging 2019: Image Processing

[San16] G. Santori, Journals should drive data reproducibility, Nature, 535(7612):355--355, 2016

[SPZ13] V. Stodden, G. Peixuan and M. Zhaokun, Toward reproducible computational research: an empirical analysis of data and code policy adoption by journals, PloS one, 8(6):e67111, 2013, Public Library of Science.

[Yaf15] B. Yaffe, Reproducibility in science, 8(371), eg5--eg5, 2015, Science Signaling.

[WDA+16] M. Wilkinson, M. Dumontier, I. Aalbersberg et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3, 160018, 2016