UNIVERSITÉ DE RENNES 1

Direction de la Recherche et de l'Innovation

Appel à projets - Année 2023

Contrats doctoraux «Ordinaires»

ECOLE DOCTORALE : matisse **POLE DE RECHERCHE : INFO**

Unité de recherche d'accueil

Unité de Recherche : IRISA

Equipe : DYLISS

Encadrement

Directrice de thèse : Emmanuelle BECKER **Unité de recherche** : IRISA

Employeur/Établissement d'origine : Univ. Rennes

HDR : • oui • ~~non~~ **Nb de thèses en cours** : 1

Mail : emmanuelle.becker@irisa.fr **Tél**. :

Co-directeur de thèse : Olivier DAMERON **Unité de recherche** : IRISA

Employeur/Établissement d'origine : Univ. Rennes

HDR : • oui • ~~non~~ **Nb de thèses en cours** : 1

Mail : olivier.dameron@irisa.fr **Tél.** :

Sujet

Titre : Knowledge-guided rules for generating context-specific views on a knowledge graph: application to biological networks

Mots clés : Data science, bioinformatics, biological networks, Semantic Web, BioPAX

Cotutelle internationale (~~oui~~/non) : **Si oui, préciser l'établissement et le pays de rattachement** :

Descriptif (15-30 lignes) :
Contexte socioéconomique et scientifique ; Les hypothèses et questions posées ; Les grandes étapes de la thèse et démarche ; Approches méthodologiques et techniques envisagées

Context :

Life science data are by nature heterogeneous and complementary. Analyzing their complex inter-dependencies requires a systemic approach based on biological networks encompassing metabolism, signaling and regulation. However, **we are facing two problems related to data science: first the data representation formalism may be different from the data analysis formalism, and second, there are multiple kinds of analysis, each requiring its specific formalism**.

Knowledge on biological networks is stored in several complementary bases. Fortunately, the BioPAX format facilitates their integration. Its semantically-rich formalism allows to reflect explicitely the complexity of the biological reality. This is an advantage as it supports subtle fine-grained analysis. However, this can be a setback at the same time, as the resulting topology may also be detrimental to some application-dependent reasoning. For example, inferring that two proteins are involved in two successive biochemical reactions requires to follow the molecular complexes composed of the first protein, the reaction catalyzed by these complexes, the molecules they produce, their components and eventually the reactions that consume them. These five intermediate steps between the two proteins are likely to have a detrimental effect on analyses based on random walks. This situation highlights the fact that a conversion step is required between the format relevant for data integration, and the format relevant for data analysis.
Likewise, the same BioPAX dataset(s) need to be converted into boolean networks for identifying controllers that drive the response of a biological systems to changes in its environment, or into guarded transition models for studying their dynamics. This situation highlights the fact that the data integration format must be dissociated from the multiple data analysis modalities.

There is a generic informatics challenge of (1) determining the data structure relevant for the desired analysis, (2) converting the original data into the appropriate formalisms, which typically relies on reasoning. Current projects typically follow ad-hoc approaches for generating the ad-hoc representations.

Objective

We hypothesize that the various reasoning-specific formalisms correspond to different abstractions of biological networks, and that generating these abstractions is a generic process guided by semantic-based rules.

This thesis aims at designing a generic knowledge-based method for generating application-dependent abstractions of semantically-rich data.

Approach

The first step will consist in making a survey of the extent the BioPAX format is used by the major biological networks databases (Reactome, KEGG, PathwayCommons, WikiPathway).

The second step will consist in formalizing the notion of biological network abstraction, and on providing different abstractions suitable for analyzing biological networks topology, controllers and dynamics.

The third step will consist in comparing the benefits of analyzing the biological networks based on these abstractions VS. based on the original BIOPAX representation.