Journal of Biomedical Informatics 56 (2015) 239–264

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

A Scientific Software Product Line for the Bioinformatics domain

Gabriella Castro B. Costa, Regina Braga *, José Maria N. David, Fernanda Campos

Federal University of Juiz de Fora – Computer Science Program, Juiz de Fora, Brazil

ARTICLE INFO

Article history:
Received 2 February 2015
Revised 4 April 2015
Accepted 19 May 2015
Available online 14 June 2015

Keywords:
Scientific workflow
Sequence alignment
Software Product Line
Ontology
Feature model

ABSTRACT

Context: Most specialized users (scientists) that use bioinformatics applications do not have suitable training on software development. Software Product Line (SPL) employs the concept of reuse considering that it is defined as a set of systems that are developed from a common set of base artifacts. In some contexts, such as in bioinformatics applications, it is advantageous to develop a collection of related software products, using SPL approach. If software products are similar enough, there is the possibility of predicting their commonalities, differences and then reuse these common features to support the development of new applications in the bioinformatics area.
Objectives: This paper presents the PL-Science approach which considers the context of SPL and ontology in order to assist scientists to define a scientific experiment, and to specify a workflow that encompasses bioinformatics applications of a given experiment. This paper also focuses on the use of ontologies to enable the use of Software Product Line in biological domains.
Method: In the context of this paper, Scientific Software Product Line (SSPL) differs from the Software Product Line due to the fact that SSPL uses an abstract scientific workflow model. This workflow is defined according to a scientific domain and using this abstract workflow model the products (scientific applications/algorithms) are instantiated.
Results: Through the use of ontology as a knowledge representation model, we can provide domain restrictions as well as add semantic aspects in order to facilitate the selection and organization of bioinformatics workflows in a Scientific Software Product Line. The use of ontologies enables not only the expression of formal restrictions but also the inferences on these restrictions, considering that a scientific domain needs a formal specification.
Conclusions: This paper presents the development of the PL-Science approach, encompassing a methodology and an infrastructure, and also presents an approach evaluation. This evaluation presents case studies in bioinformatics, which were conducted in two renowned research institutions in Brazil.

1. Introduction

Most users of scientific frameworks are scientists from a specific research field who do not have suitable training in software development. They often begin the construction of an application by copying an existing one and/or simply adapting its requirements. Software Product Line Engineering can help understand the software being developed because scientists can follow a model that specifies the product line and carefully make decisions according to their needs, for each point of variation of this model [1].

In Software Product Line (SPL) context for the scientific domain, analyzing the difficulties when specifying scientific experiments and considering the possibility of scientific applications composition, there is a need for a more appropriate semantic support for the domain analysis phase. Our hypothesis is that the use of scientific workflows [3,12] in a Software Product Line with the support of a feature model [4], associated with ontologies [5], can enable the development of experiments. As a result, this association contributes to the creation of a Scientific Software Product Line (SSPL). Considering that the tasks of identifying, tailoring and composing scientific services/algorithms/applications in scientific workflows are tedious and error prone, we propose a method with a tooled support which assists the creation of a scientific SPL, with semantic provision. In this context, this paper details the PL-Science approach, whose purpose is to support the scientists' selection process, when defining workflows, based on scientific applications, in accordance with the research requirements. Thus, through the

* Corresponding author at: Campus Universitário, S/N, UFJF, Juiz de Fora, MG, Brazil. Tel.: +55 32 2102 3378(31).
E-mail addresses: gabriellabc@gmail.com (G.C.B. Costa), regina.braga@ufjf.edu.br (R. Braga), jose.david@ufjf.edu.br (J.M.N. David), fernanda.campos@ufjf.edu.br (F. Campos).

http://dx.doi.org/10.1016/j.jbi.2015.05.014
1532-0464/

concepts of SPL, scientists can follow the models that specify the product line and make decisions according to their needs. Finally, scientists can develop a product composed of scientific applications and/or instantiated algorithms.

The PL-Science approach also has the following goals: (i) propose an architecture in order to support the implementation of a SPL for scientific applications, (ii) present an approach where the semantics is highlighted, in order to support the variability specification of the SPL, using ontologies in conjunction with feature models, (iii) implement a SSPL, evaluated by case studies in the bioinformatics area (sequencing/genetic alignment).

We can consider two main contributions of this work. The first is the development of the PL-Science approach, considering the method and the infrastructure developed. This contribution was briefly presented in [10]. The second contribution is the approach evaluation, which is discussed in this paper in depth. This evaluation presents case studies, which were conducted in two renowned research institutions in Brazil. The obtained results were able to support our hypothesis.

Some research as described in [6,4,7,17], use approaches based on ontologies to enhance SPL support when developing applications. In these studies the need to add semantic aspects in SPL variability representation is recurrent. Our work presents a way to improve SSPL domain specification using ontologies in addition to feature models, considering the scientific context and its specificities. As a result, we used the advantages of these two domain model techniques to generate scientific workflows through an SPL approach. As will be described later in this paper, we want to extract the best of both model types, i.e., the feature model will be used to support variability representation and the ontology will be used to express formal restrictions and possible inferences on these restrictions, considering that the scientific domain needs a formal specification. The 'alignment' between these models is enriched because we try to extract the semantics from both, improving the SSPL knowledge base.

The remainder of the paper is organized as follows. Section 2 describes the theoretical background. Section 3 discusses related works. An overview of PL-Science approach, with the main models, proposed architecture, and methodology is presented in Section 4. Section 5 presents the case studies. Finally, in Section 6 we present the conclusions and future works in order to improve the PL-Science approach.

2. Theoretical background

Research in Software Engineering traditionally focuses on techniques, methods and concepts that can be applied in a general context. However, Scientific Software has very specialized domains, and therefore not every technique which is applied to general software can bring good results in scientific applications [9]. This issue fosters the need for more specific research in this area, that is, studies that focus on applying techniques related to software engineering in the scientific research context.

Scientific software are fundamentally different from traditional software, mainly due to the fact that (i) there is informality in the development process of scientific software, (ii) generally, the researchers themselves and/or scientists develop the software, (iii) survey and requirements specification are both hindered because they may not appear clearly, or sometimes even be unfamiliar, in an initial research stage [10].

The importance of using scientific workflows is inherently related to how scientists currently plan a scientific experiment *in silico* and the collaboration requirements from different research centers. Because of these collaboration requirements, it is essential to organize an execution flow of scientific applications, which must be sequenced in order to perform the experiment. Thus, among the activities to be developed by scientists/researchers, we can highlight the sequencing (or composition) of programs/scientific applications, where each of these programs produces a data collection with a particular semantic and syntax. This data collection can mostly be used as input data for the next program. However, it is worth noting that program composition is not a trivial task and it can sometimes become a barrier to further analysis by researchers. One way to minimize this problem is through the use of scientific workflows, that is, *in silico* experiments are represented by means of chaining activities, and each activity is mapped to an application forming a coherent flow of information and controls. This chain of activities is called a scientific workflow [11].

According to Clements and Northrop [2] Software Product Line can be defined as "a set of software intensive systems sharing a common set of features which are managed to satisfy specific needs of a particular market segment or mission and that are developed from a common set of core assets in a prescribed way". The development process of a family of programs is divided into two phases: domain engineering and application engineering [13]. The development of core artifacts or the domain engineering phase is related to the development of reusable components from domain analysis of SPL. The process of product development, also known as application engineering, is responsible for analyzing the requirements of the application to be generated and then derive a 'concrete' product using the variability model. Finally, the generated product is available in the user's environment [3,18].

In the SPL context, variability is the ability of the system to be effectively scalable, changeable, customized or configured for use in a particular context [6]. Products incorporating variability may have advantages such as addressing various segments of the market and providing different sets of features according to their needs. A feature can be defined as a relevant system property used to capture similarities and variabilities between products of a SPL [14]. Variability models are essential for the development and management of Software Product Lines. They may contain concepts related to decisions, features or variation points, depending on the abstraction level. Through feature modeling, common and variable features of an application family are specified in order to be supported by the product line. This model should also include constraints between variation points and variants because a variation point (or a variant) can require or exclude another variation point (or a variant) [1]. However, the use of only a feature model to represent the characteristics and the restrictions in the domain is a limited resource in the scientific application context. For example, in the area of genetic sequencing/alignment feature model, it is not possible to express overall constraints and semantics required for scientific applications. These constraints are mainly related to the semantics involved in relationships between features. A higher meaning of the features can be supplied by the use of ontology as well as by the possibility to infer related concepts.

Thus, ontology technology can be used for formal representation of constraints between the variation points of a SPL. An ontology allows the use of inference mechanisms through which we can discover new knowledge. This is one of the great advantages of its use. Furthermore, the creation of restrictions in ontologies (described using OWL-DL) is simpler than the creation of statements in propositional logic.

Ontology defines a formal and explicit specification of a shared conceptualization [15]. By the use of ontologies, it is possible to establish a common understanding about objects and the relationships between them in a given domain, through a formal model [5]. In addition, the formal specification of the meaning of the terms in the ontology enables the creation of new terms by combining the existing ones [5]. Besides this, there is the possibility of using inference machines (reasoners), offering algorithms through which one