Available online at www.sciencedirect.com
ScienceDirect
Computational Biology and Chemistry
Computational Biology and Chemistry 31 (2007) 305–319
www.elsevier.com/locate/compbiolchem

Review Article

Workflow based framework for life science informatics

Abhishek Tiwari [a,*], Arvind K.T. Sekhar [b]

[a] *Informatics Division, GVK Biosciences, Hyderabad 500037, India*
[b] *School of Informatics, Indiana University (IUPUI), Indianapolis, USA*

Received 19 February 2007; received in revised form 15 June 2007; accepted 10 August 2007

Abstract

Workflow technology is a generic mechanism to integrate diverse types of available resources (databases, servers, software applications and different services) which facilitate knowledge exchange within traditionally divergent fields such as molecular biology, clinical research, computational science, physics, chemistry and statistics. Researchers can easily incorporate and access diverse, distributed tools and data to develop their own research protocols for scientific analysis. Application of workflow technology has been reported in areas like drug discovery, genomics, large-scale gene expression analysis, proteomics, and system biology. In this article, we have discussed the existing workflow systems and the trends in applications of workflow based systems.

Keywords: Workflow technology; Data pipelining; Data streaming; Visual programming; Ontology; Semantic web; Web services; Grid services

Contents

1. Introduction

In recent years, explosive amounts of biological information has been obtained and deposited in various databases. The predominant source of this "data flood" is "high-throughput" experimentation, involving simultaneous execution of hundreds or thousands of experiments. Workflow systems could become crucial for enabling scientists to deal with this data explosion. A comprehensive understanding of biological phenomena can be achieved only through the integration of all available biological information and different data analysis tools and applications. Workflow environment allows life science researchers to perform the integration themselves without involving any programming. Workflow system allows the construction of complex *in silico* experiments in the form of workflows and data pipelines. Data pipelining is a relatively simple concept. Any computational component or node has data inputs and data outputs. Data pipelining views these nodes as being connected together by 'pipes' through which data flows. In a workflow controlled data pipeline, as the data flows, it is transformed and raw data is analyzed to become information and the collected information gives rise to knowledge. The concept of workflow is not new and it has been used by many organizations, over the years, to improve productivity and increase efficiency. A workflow system is highly flexible and can accommodate any changes or updates whenever new or modified data and corresponding analytical tools become available.

* Corresponding author at: 9/11 D Bank Road Old Katra, Allahabad 211002, UP, India. Tel.: +91 9347808010.
E-mail address: abhishek.twr@gmail.com (A. Tiwari).

1476-9271/$ – see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiolchem.2007.08.009

2. Description

A workflow system (Hollingsworth, 1995) is a holistic unit that defines, manages, and executes workflow processes aided by software. The order of execution is defined by a computer representation of the workflow process logic. Internally, a workflow system uses a Workflow Language or Meta-Languages for process specification (Michael and Jörg, 1999) to define the workflow process logic, to be executed by workflow execution engine or workflow controller. Visual representation of the workflow process logic is generally carried out using a Graphical User Interface where different types of nodes (data transformation point) or software components are available for connection through edges or pipes that define the workflow process. Graphical User Interfaces provide drag and drop utility for creating an abstract workflow, also known as "visual programming". The anatomy of a workflow node or component (cf. Fig. 1) is basically defined by three parameters: (1) input metadata, (2) transformation rules, algorithms or user parameters, (3) output metadata. Nodes can be plugged together only if the output of one, previous (set of) node(s) represents the mandatory input requirements of the following node. Thus, the essential description of a node actually comprises only in- and output that are described fully in terms of data types and their semantics. Semantic integration of different data sources requires domain specific ontology (Baker et al., 1999) like Gene Ontology, Glycomics Ontology (GlycO), Proteomics Process Ontology (ProPreO), Blue Obelisk Chemoinformatics Algorithms Ontology. It is worth noting that ontology is a set of controlled vocabulary that classify concepts and define the relationships between them for the information in a specific domain which should be interpretable both by machines and humans. Ontology is also used for workflow validation without knowing applications details (input type, data type, etc.) and conversion of input data, if needed. A general workflow based framework for life sciences domain is described in Fig. 2 which has three different layers—clients layer, component and enactment layer, and database layer. The client has a Graphical User Interface for the creation of workflows along with process definition service. The user can create workflows using any combination of the available tools, services or databases in workflow system by dragging/dropping and linking graphical icons. Process definition services use Meta-Languages for Workflow and Process Modeling. In this layer workflow is rendered in workflow description language which is fallowed by workflow validation to ensure that the workflow is syntactically correct. Component and enactment layer has component services and enactment services and it can interact with database layer. Component services provide different software components, tools and services (grid and web services). An enactment service consists of one or more workflow engines in order to manage and execute workflow instances. Workflow engine or workflow controller is responsible for the part or all of the runtime control environment within an enactment service. Workflow engine has a job scheduling mechanisms which performs resources binding and component invocation. Communication or information exchange between different layers or services is normally defined in workflow definition languages or XML. Communication via XML is one of the standard ways. Database layer consist of diverse range of data sources (remote as well as local).

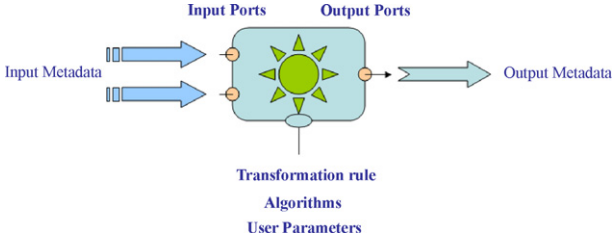

Fig. 1. The anatomy of a workflow node or component. The component properties are best described by the input metadata, output metadata and user defined parameters or transformation rules. The inputs ports can be constrained to only accept data of a specific type such as those provided by another component.

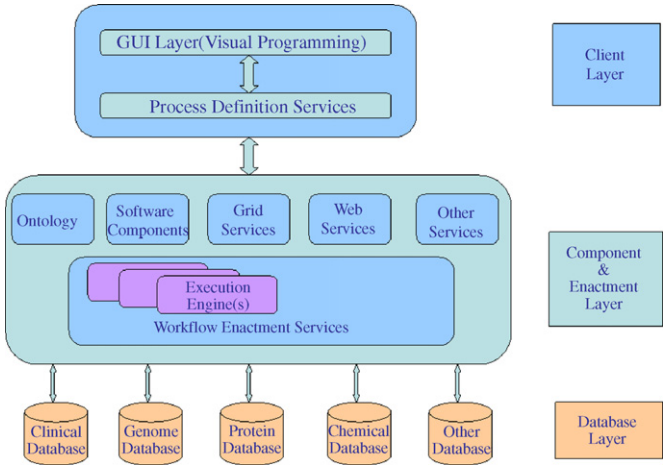

Fig. 2. A general workflow based framework for life sciences informatics.

An *in silico* analysis of the workflow is best carried in phases. In the first phase, a conceptual workflow is generated. A conceptual workflow as the name suggests is a sequential arrangement of different components that the user may require to accomplish the given task. It may quite be possible that some of the steps may in turn be composed of several sub components. The next phase converts the conceptual workflow into an abstract workflow by performing a visual drag and drop of the individual components that were figured to be a part of the workflow in the first phase. The workflow is termed abstract in that it is not yet fully functional but the actual components are in place and in the requisite order. In general, workflow systems concentrate on creation of abstract process workflows to which data can be applied when the design process is complete. In contrast, workflow systems in the life sciences domain are often based on a dataflow (Hudson et al., 2004) model, due to the data-centric and data-driven nature of many scientific analyses. Currently, there are many workflow systems available in life sciences (Table 1). These workflow systems can integrate most of the available, standard software tools (either commercial or public domain) along with different classes of programmable toolkits (Table 2).