HAL
open science

Context-oriented scientific workflow system and its application in virtual screening

Xiaoliang Fan, Patrick Brézillon, Ruisheng Zhang, Lian Li

To cite this version:

Xiaoliang Fan, Patrick Brézillon, Ruisheng Zhang, Lian Li. Context-oriented scientific workflow system and its application in virtual screening. 15th IFIP WG 8.3 International Conference on Decision Support Systems (DSS 2010), Jul 2010, Lisbon, Portugal. pp.335-345, 10.3233/978-1-60750-577-8-335 . hal-00626738

HAL Id: hal-00626738

https://hal.science/hal-00626738v1

Submitted on 27 Sep 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Context-oriented scientific workflow system and its application in virtual screening

Xiaoliang FAN[a,b,1], Patrick BRÉZILLON[b], Ruisheng ZHANG[a], and Lian LI[a,c]
[a]*School of Information Science and Engineering, Lanzhou University, Lanzhou, China*
[b] *Laboratory of Computer Sciences of Paris 6, UPMC, Paris, France*
[c] *School of Computer and Information, Hefei University of Technology, Hefei, China*

Abstract. Scientific workflow (SWF) system is gradually liberating the computational scientists from burden of data-centric operations to concentration on their decision making. However, contemporary SWF systems fail to address the variables when scientists urge to deliver new outcomes through reproduction of workflow, including not only workflow representation, but also its "context" of use. Thus, current failure is mainly due to lack of representing and managing the "context". We propose a context-oriented approach to create a SWF system more adaptive to the dynamic research environment. We present the feedbacks from interviews of multi-national virtual screening scientists, and illustrate a case study in which their decision making processes are modelled by contextual graphs as a uniform representation of knowledge, of reasoning, and of contexts. Finally, we conclude and highlight that sharing SWF intellectually with its context would make SWF as a complement to paper-based publications.

Keywords. Context, scientific workflow system, decision making, grid computing, virtual screening

1. Introduction

Scientific workflow (SWF) system is a specific workflow management system applied to science arena, and it concerns with the automation of scientific processes in which scientific tasks are orchestrated based on their control and data dependencies [1,9]. For years, SWF systems are widely applied to many applications, namely in physics, climate modeling, drug discovery process, and disaster recovery simulation [7,8].

Computational science [15], such as virtual screening [16], is an exploratory, flexible, and knowledge-intensive one that it is always re-done in an evolutionary pattern. However, contemporary SWF systems fail to address such dynamics and variables when scientists urge to deliver new outcomes through reproduction of SWF, including not only SWF representation and data provenance, but also its "context" of use. Main reason for such failure is due to the lack of representing and managing the context, defined in [12] as "which constrains something without intervening in it explicitly". For scientists, a new result modifies the context of their decision making process. Thus, it is important for them to accounting for context. If not handling the

[1] Corresponding Author: Xiaoliang FAN, LIP6, UPMC, 104 Avenue du Président Kennedy, 75016, Paris, France; E-mail: xiaoliang.fan@gmail.com

context during the scientific discovery, SWF system could not shift itself to a practice to address the specific focus and context of the scientist.

We propose a context-oriented SWF system along a user-centered approach. Context can be associated with SWF in three main ways. Firstly, in the usual way, there is a library of SWFs, and context acts as an interface (middleware, context-aware system, etc.) to determine which SWF should be chosen from the library under a specific focus of the user. In the second way, context supports the assembling of components of the SWF, which must be recompiled each time. Thirdly, context is expressed by context graphs (CxGs), combined with SWF representation according to contextual elements (CEs), which will enrich the SWF representation, enhance its flexibility, and help to make a better decision for scientists.

Making "context" explicit in SWF system would formalize scientists' research, strategies, and customization information, where elements of knowledge, of reasoning and of contexts are represented in a uniform way. We describe series of feedbacks from interviews of multi-national virtual screening scientists, and illustrate a case study in virtual screening. We generalize groups of contextualized information and model them by contextual graphs. Finally, we conclude that the potentials of intelligently sharing workflow with its context would make scientific workflow as a complementary of paper-based publications.

Hereafter, the paper is organized in the following way. Section 2 presents the problem addressed in scientific discovery and why current SWF systems fail to address the dynamic challenges. Section 3 discusses and comments briefly previous works on context-based applications in order to point out what are reusable while problems remain to solve for SWF flexibility. We then propose our context-orient SWF solution in Section 4. Section 5 presents a case study in virtual screening. The general conclusion and future work in Section 6 closes the paper.

2. Problem Addressed

We are in the era of data-centric scientific research, in which hypotheses are more and more generated by combining and mining the pool of data already available [13]. However, current SWF systems fail to address such challenge for the following reasons:

- *Lack of "Context"*. Current SWF systems are not capable of making the "context" of the case (including task at hand, environment, user's mood, etc.) accessible to scientists. SWF shifts scientific procedural, but without representation and management of context, the procedural could not shift itself to a specific practice to address current focus of the scientist.
- *Inflexible decomposition of workflow representation*. SWF is not atomic for the user, who needs to carry out the actual work at a much more "fine-grained" level, such as "changing the parameters whenever needed in real-time".
- *Push oriented perspective*. Most SWF systems force scientists to do what they are obliged to do rather than what they are able to do in a unique context. Such a "push" way gives rise to inflexible practices, and blocks the decision-making process.
- *Sharing of SWF representation is far from enough*. Smart and intelligent rerun of SWF automatically is urged by scientists to deliver new outcomes when fresh data with new context become available [13].

3. Related Works

Context has been playing an important role in a number of decision making domains for a long time. Context-based intelligent assistant system, a new generation of system which combines the strengths of knowledge based system with those of intelligent decision-making assistant system, has successful applications in diagnosis of equipment for EDF, French national power company [2], incident management on the subway [3], contextualization of a social network [4,5], medical image retrieval [6], etc.

The need for flexibility and dynamicity in scientific workflow have long been recognized as a critical issue to adapt to changing circumstances of scientific discovery. A variety of artificial intelligence techniques have been brought to bear for intelligent SWF representation and automation, while guaranteeing data quality:

- [18] discusses the need to assist scientists at a higher level that requires capturing and exploiting scientific knowledge about the software and data used in computational experimentation. WINGS [19] also presents the idea of reasoning with SWF at the knowledge level using ontologies and rules.
- myExperiment [9] enables social networking around SWFs in bioinformatics applications, builds up a SWF repositories, and provides community support for social tagging, comments, ratings and recommendations by adopting content-sharing tools and other Web 2.0 technologies.
- DECLARE [17] identifies three flexibility mechanisms for workflow, which could suspend the decision makings, change or ignore the workflow representation separately when needed in a flexible way.
- MOTEUR [20] presents a novel scientific workflow system combining an user-friendly workflow representation with flexible and efficient execution strategies, such as enabling dynamic extension of the data sets.

However, it is more and more proven that the general approach doesn't work well, because of not closely working with the user. A context-oriented approach will not only make use of previous success of context-based intelligent assistant systems, such as capturing contextual information through context graphs, but also work closely with scientists to promote intelligently share of SWFs through the process of incremental knowledge acquisition. One contribution also embraces making available a tool for teaching virtual screening because students (Master and Ph.D) will know which methods can be used at one step and the rationale (i.e. the contextual elements) for the choice of any methods. Further more, this work has been carried out in an international-level cooperation with virtual screening scientists from both France and China.

4. Proposed Context-oriented Approach

We propose a context-oriented solution and believe that making “context” explicit in scientific workflow system could promote a SWF more adaptive to the dynamic environment and enhance its intelligence to facilitate an effective decision-making. This results following key features of context-oriented SWF system:

- “Context” is the essential element for each scientist to represent his/her own “practice” (which is considered as the contextualization of “procedure”) in a SWF, to address a focus in a given context.

- An "experience-based knowledge base" is designed and built, and the knowledge base is enriched and modified through a process of incremental knowledge acquisition.
- An interface is developed for acquisition and management of the "experience-based knowledge base" .
- The solution should be a user-centered one, that less "push" but more "participating" and "interactive" oriented SWF is represented.

Applying scientific workflow to improve decision making outcomes in virtual screening is a very difficult issue. On one hand, workflow represents a compilation of well-established procedures, robust and usable in a number of domains. On the other hand, virtual screening is a domain where there is no unique solution but a very large number of solutions that depend on a number of factors specific of products, molecule, situation, software, environment, etc. Representing these factors as contextual elements, then assembling contextual elements as contextual graphs, seems to be the first priority for approaching a complete representation of SWF and make available an experience-based knowledge base. Our recent achievement is mainly presented in a case study.

5. Case Study

5.1. Background

Virtual screening scientist published their finding on crystal structure of "Polymerase PA_C-$PB1_N$ complex from an avian influenza H5N1 virus" in *Nature* [10,11], considered as the first successful result of drug discovery research on avian influenza H5N1 virus. "PA_C" is a protein, which is a kind of H5N1 virus, and its toxicity will only take effect when combining with a small molecule called "$PB1_N$". In general, this research aims to find dozens of molecules from millions of molecules in ZINC database [21], which could best block the binding between PA_C and $PB1_N$. As a result, the viral replication of H5N1 is abolished, hereafter "dozens of molecules" become candidates for anti-H5N1 drug. Based on this finding, we are running a project entitled "Virtual screening research on polymerase PA_C", which aims to dock [22] about 7.7 million small molecules separately on PA_C. The application is not only a time-consuming one in which millions of computings are expected to perform by docking software (such as Dock 6.0), but also a very flexible one that there is no unique solution for each computings because they vary from each other on selecting docking software.

In order to address such flexibility challenges, we believe the starting point should be a representation of scientists' workflow in formalism where elements of knowledge, reasoning and contexts are represented in a uniform way. Hence, we firstly generalize feedbacks on procedurals and decision making processes of virtual screening research, through questionnaires to scientists in Section 5.2. Then we make interviews to scientists on detailed information during their research, through which contextualized information, or known as contextual elements (CEs), are obtained in Section 5.3. Further more, we begin to use context graphs (CxGs) to model the CEs and preliminary result is presented in Section 5.4.

5.2. Feedbacks from Virtual Screening Scientists

Through questionnaires to several scientists, we get the knowledge that there are mainly two rounds of screenings for virtual screening research:

- First round screening aims to perform rigid docking between millions of molecules and the protein separately.
- Second round screening selects higher-score molecules (about 10% of total molecules) after the first round screening, and performs flexible docking between selected molecules and protein separately.

More in detailed, we present the general steps, including decision makings process, of "protein preparation", which is a very important portion of virtual screening during the first round screening:

Step 1. Name the targeted protein (PA_C) and get its structure.
Step 2. Execute the necessary preparation works:
Step 2.1 remove unrelated molecular **(**such as water);
Step 2.2 add hydrogen and charge;
Step 2.3 optimize the protein (PA_C), etc.
Step 3. Locate the binding site of protein (PA_C) iteratively:
Step 3.1

- Perform the docking calculation between the protein (PA_C) and $PB1_N$,
- Get its RMSD value (RMSD_0, in chemoinformatics, RMSD is a measure of the distance between a crystal structure conformation and a docking result),
- Compare the conformation of $PB1_N$ before the docking, with the new one after the docking;

Step 3.2

- Do the first round virtual screening of protein (PA_C) with 1000 ligands,
- Get the RMSD value separately (RMSD_1, … , RMSD_1000);

Step 3.3 compare the value of RMSD_0 with RMSD_1, …, RMSD_1000;
Step 3.4 Is the result good?

- If "YES" (two indicators: conformations of two $PB1_N$ mentioned in *Step 3.1* are similar, and the difference between the value of RMSD_0 and Minimum{RMSD_1, ..., RMSD_1000} is not huge), then choose this binding site as the one which will be used in the large-scale docking process later (with million of ligands).
- if NOT, re-configure the parameters and re-do *Step 3.1, Step 3.2, Step 3.3* (iteratively if necessary), until the answer of *Step 3.4* is "YES".

5.3. Results Generalized through Interviews

Based on the feedbacks from scientists, we make interviews on detailed information during their research, through which two results are generalized:

1st result: Virtual screening is an adaptive research, and computing is always re-done in a repeatable and evolutionary pattern. There are two aspects for the 1st result:

- **Scientists always ignore or add a portion of SWF when re-executing the SWF.** For example, during the first-round screening, scientists firstly do the "protein preparation" for PA_C. But during the second-round, there is no need to repeat such step for the new SWF, because the same protein will be performed. Another example, second-round screening would be flexible

docking. Thus, they add a new step in the second-round called "protein optimization".

- **Modifying the parameters is a necessity when performing a new-round of screening.** For example, in the second-round, they change the parameter from "flexible = NO" to "flexible = YES", because it is the time to do a precise and flexible screening. The reason behind this sort of modification, is mainly due to updating the research objectives, such as computing budget, how much time they have, etc.

2nd result: Sharing of SWF means not only to share of representation, but also to share of "context". There are two aspects for the 2nd result:

- **It is necessary to distinguish virtual screening computings under various "contexts".** Once, we preform two computings, one to local computer in Lanzhou University, the other to a distributed computer in Tsinghua University in Beijing. We thought two computings should be finished at the same time. But it turned out that the latter one is finished 3 weeks after the former one, because the latter computer has only one CPU while the former one is a cluster with 64 CPUs. Similar accident happened when working on different version of software (Dock 5.0 works differently from Dock 6.0). Obviously, such lessons learnt indicates that the "context" of use is missing.
- **"Context" plays a very important role in two ways of SWF sharing.** Firstly, sharing of SWF with context is considered as a complementary part of the paper-based publication. Other scientists would evaluate your hypothesis in a much more precise way by re-executing the SWF when the context is met. In second way, it enables the reproduction, transformation, and evolution of other's "old" SWF to your brand "new" one. Take virtual screening on H5N1 virus for example, since H5N1 virus shares a general structure with H1N1 virus (95% similarity, through the protein sequence analysis of RNA polymerase PAc between H5N1 and H1N1), it is possible to share SWF representation of H5N1 virtual screening with H1N1 researchers.

These results generated through interviews, is vital to the extraction of contextual information as contextual graphs, which promotes a user-centered approach of decision-makings for virtual screening scientists.

5.4. Modeling SWF with Contextual Graphs

The piece of software Contextual Graphs (CxGs) is presented, explained and available at www.cxg.fr. A contextual graph represents the different ways to solve a problem. It is a directed graph, acyclic with one input and one output and a general structure of spindle. A path in a contextual graph corresponds to a specific way (i.e. a practice) for the problem solving, represented by the contextual graph. It is composed of elements of reasoning and of contexts, the latter being instantiated on the path followed (i.e. the values of the contextual elements are required for selecting a branch, i.e. an element of reasoning among several ones). Figure 1 provides the definition of the elements in a contextual graph (actions, contextual elements, sub-graphs, activities and temporal branching). A more complete presentation of this formalism and its implementation can be found in [5].

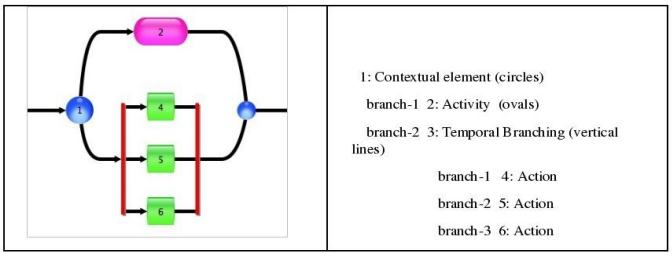

Figure 1. Elements of a contextual graph

We begin to use CxGs to model our case study in virtual screening. Generally speaking, as Section 5.2 indicated, there are two rounds of virtual screening. However, the exception exists, which is demonstrated in Figure 2. By interview, scientist first proposes a question whether there are less than 100 thousands molecules (1). If so, scientist could ignore the first phase of rigid docking and directly move to the flexible docking (2). Else, there are more than 100 thousands molecules, it is necessary to do the two rounds of screening separately (3, and 4). Finally, when the screening is finished, scientist will analyze the result (5). In our case study, scientist has 7.7 millions molecules, so the path of "1→3→4→5" is followed.

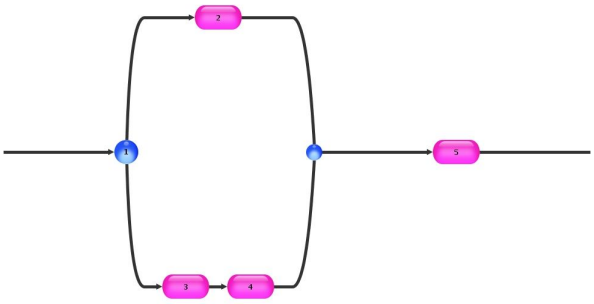

Figure 2. Contextual graph "The process of virtual screening"

1 : What is your screening scalability?
less than 100 thousands molecules 2 : Activity: direct flexible screening
more than 100 thousands 3 : Activity: first-round rigid screening
4 : Activity: second-round flexible screening
5 : Activity: analyze the results and virtual screening is finished

Figure 3 shows what happens inside "first-round rigid screening" (Activity 3 in Figure 2). In our case, scientist performs first-round screening between each molecule and Polymerase PA_C, separately. Firstly, check if it is a rigid screening (2). If "No", quit for checking again (3); if "Yes" then do the following: prepare the protein (1), then locate the binding site (7), and perform the docking (6). Later, by analyzing the result of first-round rigid screening (4), select the top-scorer molecules and store them for second-round screening (5).

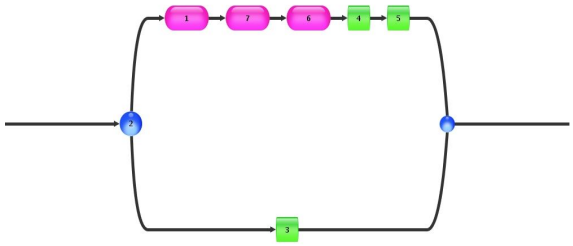

Figure 3. Contextual graph "Activity: first-round rigid screening" (Activity 3 in Figure 2)

2 : Is it a rigid screening?
- Yes — 1 : Activity: protein preparation
  - 7 : Activity: locate the binding site
  - 6 : Activity: perform the docking
  - 4 : analyze the first-round rigid screening
  - 5 : select the top-scorer molecules, and store them for second-round flexible screening
- No — 3 : quit this screening until "rigid screening is ready"

Figure 4 shows the comparison between SWF representation of "protein preparation" without introduction of context (Left), and contextual graph of "protein preparation" (Right, Activity 1 in Figure 3).

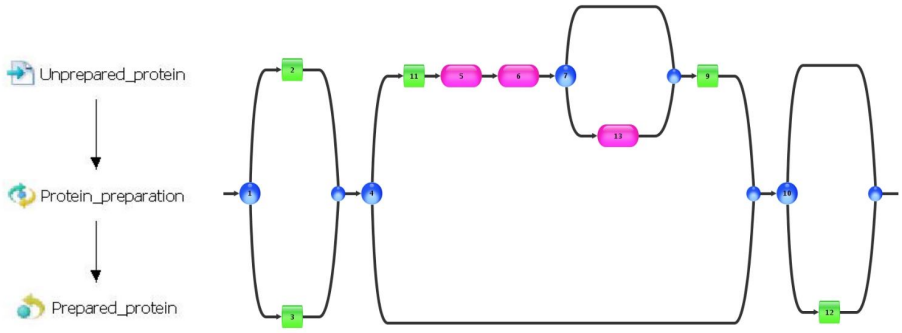

Figure 4. (Left) SWF representation of "protein preparation" without introduction of context

(Right) Contextual graph "Activity: protein preparation" (Activity 1 in Figure 3)

1 : Can you find the protein by yourself?
- Yes — 2 : Download it from "Protein Data Bank"
- No — 3 : ask for help from your colleagues until you get the protein

4 : Do you need to do "protein preparation"?
- Yes — 11 : Enter series of parameters during "protein preparation"
  - 5 : Activity: remove unrelated molecules
  - 6 : Activity: add hydrogen and charge
  - 7 : Is it a rigid or flexible screening?
    - rigid
    - flexible — 13 : Activity: optimize the protein
  - 9 : store the protein prepared in the database
- No

10 : Are you satisfied with the result of "protein preparation"?
- Yes
- No — 12 : Modify the series of parameters and redo the "protein preparation"

As a result, from Figure 2 to Figure 3, then to Figure 4 (right), contextual graphs in different granularities work together to represent the virtual screening process. Further more, through the comparison in Figure 4, we could easily conclude that contextual graphs provide a uniform representation of elements of knowledge, of reasoning, and of contexts, in which SWF dynamic and flexibility is expressed in a clear way:

- Before the contextual graphs was corresponding to SWF representation (Left picture of Figure 4), SWF is rigid as several pieces of components assembled with sequence of input data (Unprepared_protein), workflow component (Protein_preparation), and output data (Prepared_protein). However, the virtual screening research is a flexible one that scientists could not afford to re-design the representation every time when trivial changes (caused by the change of context) is urged for a reproduced SWF. By introducing contextual graphs (Right picture of Figure 4), scientists could easily model flexibilities and decision-makings in contextual elements, which would control the assembly of ignoring or adding a portion of SWF whenever needed.
- Further more, when CxGs are included in the workflow compilation, the value of contextual element will lead to invoking a certain workflow component, which enables multi-outputs of decision-making process. For example, such outputs of whether to invoke Activity 13 (right picture of Figure 4) or not, is manipulated by Contextual element 7.

The power of contextual graph has not yet come (at least not directly) in the designing phase of context-oriented SWF systems, but it is considered as an important and useful tool to assist the user in reacting quickly when facing with unexpected circumstances during the decision making process. To enhance this function, it is expected to immigrate the merits of contextual graphs with current scripts of SWF representation (such as BPEL [9]).

6. Conclusion and Future Work

Virtual screening research is an adaptive research, and decisions are always re-made in a repeatable and evolutionary pattern. Unfortunately, current scientific workflow systems fail to adapt such dynamicity and variables.

The potential benefits of making “context” explicit in SWF system will:

- represent and formalize virtual screening scientists' idea, strategy, knowledge, and reasoning in a uniform way;
- encourage “sharing of SWF and data under a specific context” within scientific communities, as a complementary form of research outcome, which should be curated and archived along with paper-based publications (see recent *Nature*'s Editorials on “Data sharing” [14], and huge success of myExperiment project [19]).

Context-oriented SWF open the door to significant new capabilities for automated scientific discovery and reasoning over data, knowledge, and discovery. Applying scientific workflow to virtual screening is a very difficult attempt. Representing dynamicity and flexibility as contextual elements, then assembling contextual elements as contextual graphs, seems to be the first priority for approaching a complete representation of SWF. The association of such a context model with scientific

workflows, would lead to an efficient use of a robust method in virtual screening once not formalized at all. As soon as we get enough contextualized information during virtual screening, next step would use contextual graphs to build up an experience-based knowledge base for the context-oriented SWF systems. It is also vital to get more and more computational scientists (such as virtual screening scientists) and distributed computing scientists (such as grid/cloud computing) involved in this inter-discipline area to promote collaborations in a shared context. We believe that, "data-centric" and "knowledge-intensive" are not competitive but complementary for accelerating the pace of scientific discovery, such as virtual screening.

Acknowledgment

This work is supported by National Natural Science Foundation of China (90912003, 60773108, 90812001), Ministry of Science and Technology of China (2005DKA64001), and China Scholarship Council (2008618047).

References

[1] Altintas, I., Berkley, C., Jaeger, E., Jones, M., Ludascher, B., Mock, S., Kepler: an extensible system for design and execution of scientific workflows, *In proceedings of 16th International Conference on Scientific and Statistical Database Management*, 423-424, 2004.

[2] Brézillon P, Bau D-Y, Diagnostic basé sur un modèle des systèmes de contrôle dans les postes des réseaux d'énergie: l'expérience SEPT, *Revue d'Intelligence Artificielle*, **6(4)** (1993), 407-430.

[3] Brézillon P., Cavalcanti, M., Naveiro, R., Pomerol J-Ch., SART: An intelligent assistant for subway control, *Pesquisa Operacional*, *Brazilian Operations Research Society*, **20(2)** (2000), 247-268.

[4] Brézillon P., Context and virtual communities in a firm, *Computing And Informatics*, 23(2004), 115-131.

[5] Brézillon P., Task-realization models in contextual graphs. In Dey, A.K., Kokinov, B., Leake, D., Turner, R., eds.: Modeling and Using Context: 5th International and Interdisciplinary Conference (CONTEXT'05), Springer Verlag, **LNAI 3554** (2005), 55–68.

[6] Brézillon, P., Daniel R., A Context Model for Content Based Medical Image Retrieval, *Medical Imaging Technology,* **25(5)** (2007), 327-332.

[7] Yu, J., Buyya, R., A Taxonomy of Scientific Workflow Systems for Grid Computing. *In special issue on Scientific Workflows, SIGMOD Record 2005*; **34(3)** (2005), 44-49.

[8] Huajian Zhang, Xiaoliang FAN, Ruisheng Zhang, et al, Extending BPEL2.0 for Grid-Based Scientific Workflow Systems, *In proceedings of IEEE Asia-Pacific Services Computing Conference (IEEE APSCC'08)*, Yilan, Chinese Taiwan, 757-762, 2008.

[9] CA Goble, D De Roure, myExperiment: social networking for workflow-using e-scientists, *In proceedings of 2nd workshop on Workflows in support of large-scale science (WORKS'07)*, Monterey, USA, 1-2, 2007.

[10] He X., Zhou J., Bartlam M., Zhang R., Ma J., et al, Crystal structure of the polymerase PA(C)-PB1(N) complex from an avian influenza H5N1 virus. *Nature*, **454** (2008), 1123–1126.

[11] Yuan P., Bartlam M., Lou Z., Chen S., Zhou J., et al, Crystal structure of an avian influenza polymerase PAN reveals an endonuclease active site, *Nature*, **458** (2009), 909-913.

[12] Brézillon P., Characteristics of context. *In Proceedings of the 19th International In: M. Ali & R. Dapoigny (IEA/AIE'06)*, Springer Verlag, LNAI 4031, 146-154, 2006.

[13] Tony hey, Stewart tansley, and Kristin tolle, *The Fourth Paradigm: Data-Intensive Scientific Discovery*, Microsoft research, REDMOND, WASHINGTON, 137-145, 2009.

[14] Paul N.S., et al, Post-publication sharing of data and tools, *Nature*, **461** (2009), 171-173.

[15] P.J. Roache, *Verification and validation in computational science and engineering*, Hermosa Publishers, Albuquerque, NM, 1998.

[16] Kitchen D. B., Decornez H., Furr J. R., Bajorth J., Docking and scoring in virtual screening for drug discovery: Methods andapplications. *Nat. ReV. Drug DiscoVery,* **3** (2004), 935-949.

[17] W.M.P. van der Aalst, M. Pesic and H. Schonenberg, Declarative workflows: Balancing between flexibility and support, *Computer Science - Research and Development*, **23(2)** (2009), 99–113.

[18] Yolanda Gil, From data to knowledge to discoveries: Artificial intelligence and scientific workflows, *Scientific Programming*, **17(3)** (2009), 231-246.

[19] Yolanda Gil, et al, Wings for Pegasus: Creating Large-Scale Scientific Applications Using Semantic Representations of Computational Workflows, *In proceedings of the 19th Annual Conference on Innovative Applications of Artificial Intelligence (IAAI'07)*, Vancouver, British Columbia, Canada, 1767-1774, 2007.

[20] T. Glatard, J. Montagnat, D. Lingrand, and X. Pennec, Flexible and efficient workflow deployement of data-intensive applications on grids with MOTEUR, *International Journal of High Performance Computing and Applications*, **22** (2008), 347-360.

[21] Irwin and Shoichet, ZINC:A Free Database of Commercially Available Compounds for Virtual Screening, *J. Chem. Inf. Model.,* **45(1)** (2005), 177-182.

[22] Chen Y., Shoichet BK., Molecular docking and ligand specificity in fragment-based inhibitor discovery, *Nature Chemical Biology,* **5 (5)** (2009), 358-364.