BioSciCumulus: um portal para análise de dados de proveniência em *workflows* de biologia computacional

Débora Pina[1], Vinícius Campos[1], Vítor Silva[1], Kary Ocaña[2], Daniel de Oliveira[3], Marta Mattoso[1]

[1]COPPE/Universidade Federal do Rio de Janeiro (UFRJ), Brasil

[2]Laboratório Nacional de Computação Científica (LNCC), Brasil

[3]Instituto de Computação - Universidade Federal Fluminense (IC/UFF), Brasil

{deborabpina, vinicius.s.campos}@poli.ufrj.br, karyann@lncc.br, danielcmo@ic.uff.br, {silva, marta}@cos.ufrj.br

Resumo. *A gerência de experimentos científicos tem sido facilitada por meio de sistemas de workflows científicos (SWC). No entanto, a análise dos resultados ainda encontra dificuldades devido ao volume e a heterogeneidade dos dados gerados. Para auxiliar a análise dos experimentos, os SWC capturam dados de proveniência que rastreiam os dados da execução do workflow. Ainda assim, a análise por parte do usuário esbarra na dificuldade de conhecimento da linguagem de consultas e da modelagem dos dados de proveniência para realizar a análise. Para apoiar essas questões, este artigo propõe o Portal BioSciCumulus para facilitar a submissão de workflows científicos no domínio da bioinformática em ambientes de Processamento de Alto Desempenho (PAD) e a análise de dados, sem a necessidade de o usuário configurar o ambiente de PAD ou especificar as análises via sintaxe de linguagens de consulta.*

Abstract. *The management of scientific experiments has been supported by Scientific Workflow Systems (SWS). However, result data analysis still presents difficulties due to the volume and heterogeneity of data generated. To assist in the experiment analysis, SWS capture provenance data that track workflow execution data. Nevertheless, the analysis task may be not simple since it requires user expertise in query languages and the modeling of the provenance data to carry out the analysis. To support these issues, this paper proposes the BioSciCumulus Portal to facilitate scientific workflow submission in the bioinformatics domain in High Performance Computing (HPC) environments and data analysis, without the need for the user to configure the HPC environment or to specify their analyses via query language syntax.*

1. Introdução

O Brasil detém uma vasta área marinha, denominada como a Amazônia Azul[1], onde habitam organismos de ocorrência restrita ao Brasil. Muitos destes organismos ainda não são conhecidos, e são uma fonte potencial para a descoberta de novas informações sobre genes, relativos aos organismos marinhos presentes na área, que podem levar ao desenvolvimento de novas drogas ou biomarcadores (Andrade *et al*. 2017). Nesse

[1] https://www.marinha.mil.br/content/amazonia-azul-0

sentido, a Rede Nacional de Pesquisa em Biotecnologia Marinha[2] (BiotecMar) realiza pesquisas marinhas de biodiversidade e prospecção em nível especializado, o que permite ao Brasil se posicionar como um potencial produtor em tecnologia, processos e matéria prima. A BiotecMar possibilita tanto o sequenciamento das amostras brasileiras como as análises "ômicas", *i.e.*, genômica, metagenômica, *etc*. (Simon e Daniel 2011). Esse tipo de análise depende de simulações computacionais complexas para o processamento de grande volume de dados.

Tais simulações são compostas pelo encadeamento de programas científicos e podem ser representadas pela abstração de *workflows* científicos (Davidson e Freire 2008). Como esses *workflows* processam um grande volume de dados, são preferencialmente executados em ambientes de Processamento de Alto Desempenho (PAD). Entretanto, o grande volume de dados produzidos por um único *workflow* de metagenômica e metatranscriptômica da biologia marinha, por exemplo, dificulta as análises por parte dos usuários. Muitos dos Sistemas de *Workflows* Científicos (SWC) existentes, como o Pegasus (McLennan *et al*. 2015) e o Swift/T (Wozniak *et al*. 2013), gerenciam tanto a execução do *workflow* nestes ambientes quanto capturam os dados de proveniência relativos ao *workflow* (Davidson e Freire 2008) (que correspondem ao histórico dos dados relacionados à estrutura, execução e origem dos dados manipulados). Além dos dados de proveniência "tradicionais", os usuários também necessitam de dados específicos do domínio (conteúdo de arquivos produzidos e consumidos pelo *workflow*), que são essenciais para investigar comportamentos e confirmar (ou refutar) hipóteses científicas.

Mesmo com um SWC que integra, em um mesmo repositório, os dados de proveniência, de sua execução e de domínio associados ao *workflow*, nem sempre é simples definir consultas para permitir a análise de dados. Nestes SWC, os usuários são capazes de submeter consultas, porém é necessário usar linguagens declarativas como o SQL (*Structured Query Language*), SPARQL ou Prolog. Isto requer um esforço de aprendizado por parte do usuário, e tal conhecimento não pode ser exigido dos usuários para que os mesmos possam realizar as suas análises (Gesing *et al*. 2017). De forma a mitigar tal problema, portais científicos têm sido propostos para auxiliar a modelagem, o monitoramento e a análise de dados gerados por *workflows* científicos (Gesing *et al*. 2017). Tais portais integram em um mesmo ambiente uma interface para modelagem e submissão de *workflows*, um SWC para a execução de *workflows*, ferramentas de monitoramento da execução e consulta aos dados produzidos. No quesito consulta aos resultados obtidos de *workflows*, portais proveem funcionalidades para a visualização dos dados e a submissão de consultas "pré-programadas", *i.e.*, o usuário não pode adicionar/editar consultas ou investigar outros dados de interesse. Além disso, a grande maioria dos SWC exige o término da execução dos *workflows* para disponibilizar os dados para as análises requisitadas pelos usuários (Mattoso *et al*. 2015), sendo assim, portais com esses SWC possuem recursos limitados de processamento de consultas durante a execução.

Diante deste cenário, este artigo propõe um portal, o BioSciCumulus, cujos objetivos são apoiar a submissão de *workflows* e gerenciar a interação com o usuário na execução de *workflows* científicos de bioinformática e biologia computacional em

[2] http://biotecmar.com.br

ambientes de PAD, assim como possibilitar as análises dos resultados obtidos por meio de consultas. O BioSciCumulus é voltado para atender aos requisitos, levantados em Gesing *et al*. (2017), de modelagem de *workflows* científicos por meio de recursos gráficos, de monitoramento da execução de *workflows*, de apoio à gerência da infraestrutura computacional em ambientes de PAD e de análise de dados científicos. Este portal adota o SciCumulus (Oliveira *et al*. 2010) como seu SWC com capacidade de PAD. Apesar de o SciCumulus permitir consultas *ad-hoc* ao longo da execução do *workflow*, a interface textual disponível dificulta essa interação por parte de um usuário não familiarizado com computação, em especial na definição de *workflows* e consultas. Além de prover a possibilidade de usuários executarem as consultas específicas para o seu experimento, o BioSciCumulus também auxilia os usuários a definir a consulta e analisar seu resultado, sem precisar conhecer a sintaxe de uma linguagem de consultas. Do ponto de vista de aplicações, o BioSciCumulus foi instanciado para o domínio da biologia marinha computacional propiciando a execução e a análise dos *workflows* SciPhy (Ocaña *et al*. 2011b), SciEvol (Ocaña *et al*. 2012), SciHmm (Ocaña *et al*. 2011a) e SciMG (Benza *et al*. 2015), sendo esses importantes no contexto da rede BiotecMar.

Este artigo está organizado em três seções, além dessa introdução. A Seção 2 discute os trabalhos relacionados. A Seção 3 apresenta o portal BioSciCumulus mostrando a submissão de *workflows* e a análise de dados ao utilizar experimentos de biologia marinha computacional. Finalmente, a Seção 4 conclui este artigo.

2. Trabalhos Relacionados

Existem diversos trabalhos na literatura que propõem portais para apoiar a modelagem de *workflows* científicos e a análise de dados (Abouelhoda *et al*. 2012, Gesing *et al*. 2017, McLennan *et al*. 2015, Nguyen *et al*. 2015). O HubZero (McLennan *et al*. 2015) agrupa no portal um conjunto de comandos que interagem com o SWC Pegasus para definir a estrutura do *workflow* a ser executado, gerar um plano de execução do *workflow* e, inclusive, monitorar a execução em ambientes de PAD. No que diz respeito à análise de dados, o HubZero permite processar somente consultas após o término do *workflow*, impossibilitando assim a análise em tempo de execução. Os demais portais, incluindo a cobertura feita em Gesing *et al*. (2017), possuem consultas predefinidas, sem a especificação de parâmetros para o acesso aos dados de domínio.

O Tavaxy (Abouelhoda *et al*. 2012) é um portal voltado ao domínio da bioinformática, que permite a submissão de *workflows* científicos nos SWC Taverna e Galaxy. O Tavaxy é capaz de gerenciar chamadas remotas de aplicações por meio de serviços *Web* (utilizando o Taverna) e chamadas em infraestruturas locais (utilizando o Galaxy). Entretanto, o Tavaxy permite apenas consultas fixas, que consideram alguns valores de atributos em um mesmo conjunto de dados, ou baseadas em análises globais, como o tempo de execução do *workflow*. O portal WorkWays (Nguyen *et al*. 2015) é um portal semelhante ao Tavaxy, sendo que ele apoia intervenções em tempo de execução, não é específico a um domínio e é baseado no SWC Kepler.

Diferentemente, o FireWorks (Jain *et al*. 2015) e o OpenMOLE (Reuillon *et al*. 2013) correspondem a portais que apoiam a gerência de *workflows* científicos com a captura de dados de proveniência, favorecendo análises baseadas nos rastros de proveniência gerados a partir da execução de experimentos científicos. Além disso, semelhante ao WorkWays, tais portais apoiam intervenções durante a execução de

workflows científicos. Assim, os usuários são capazes de decidir por ajustes na especificação do *workflow* ou mesmo nos dados de entrada. Todavia, tanto o FireWorks quanto o OpenMOLE se restringem a consultas pré-definidas. Diferentemente, o Portal BioScicumulus se destaca ao oferecer consultas interativas aos dados de proveniência e de domínio ao longo da execução do *workflow* em ambientes de PAD via SciCumulus.

3. Portal BioSciCumulus

Para apoiar a submissão e a análise de *workflows* científicos no domínio de biologia marinha, o Portal BioSciCumulus, baseado no SWC SciCumulus (Oliveira *et al*. 2010) e nas suas extensões (Silva *et al*. 2014), segue a arquitetura apresentada na Figura 1, que consiste em quatro componentes (representados em retângulos azul-escuros): *Interface Gráfica*, *Configuração de Workflows*, *Submissão de Workflows* e *Processamento de Consultas*. O componente *Interface Gráfica* consiste em uma aplicação Java *Web* com *Servlet*, que recebe e processa requisições nos componentes de configuração e submissão *workflows*, assim como no de processamento de consultas. Para o desenvolvimento das páginas *Web*, utilizou-se a tecnologia JSP (*Java Server Pages*), além de arquivos no formato HTML, CSS e JavaScript.

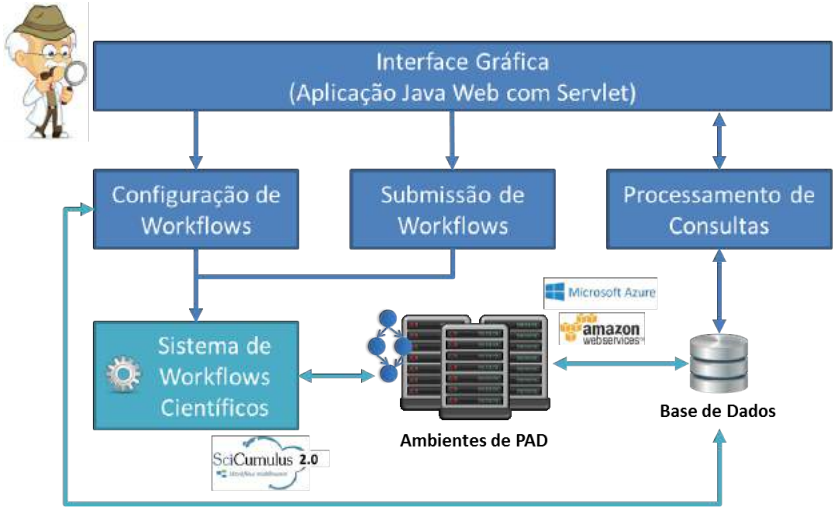

Figura 1: Arquitetura do Portal BioSciCumulus

Uma vez que a interface processa a requisição, ela é enviada ao componente de *Configuração de Workflows.* O componente de *Configuração* recebe da *Interface* um arquivo XML (arquivo de configuração do SciCumulus, que descreve as atividades do *workflow* – invocação de programas – e as relações de dados consumidos e produzidos em cada atividade) com a especificação do *workflow* e os programas a serem executados. Ele registra na base de dados a especificação do XML e configura os diretórios para futuras execuções. Dentre as informações registradas na base de dados, estão o rótulo (*tag*) do *workflow*, que também é usado para o nome da base de dados e um arquivo compactado (em formato *zip*) que contém os arquivos referentes aos programas científicos invocados nas atividades.

No componente de *Submissão* de *Workflow*, o usuário solicita a execução paralela de *workflows* já cadastrados no Portal. Nesse caso, o BioSciCumulus gerencia cada submissão criando uma instância do *workflow* científico e executando-a em um ambiente de PAD disponível (*e.g.,* um *cluster* ou nuvem). Cada *workflow* configurado previamente apresenta uma *tag* específica (que foi definida pelo usuário). Como um

mesmo *workflow* pode ser executado diversas vezes, o BioSciCumulus gera um identificador para cada execução, chamado de rótulo de execução do *workflow* (*exectag*). Para gerar os diferentes *exectags*, o BioSciCumulus realiza uma consulta à tabela *eworkflow* (que armazena os dados dos *workflows* gerenciados pelo BioSciCumulus) e define o próximo rótulo sem duplicidade.

Ainda em relação ao componente de *Submissão*, o usuário não necessita informar valores para os parâmetros que definem o diretório do *workflow*, o nome da base, seu usuário e senha, uma vez que o portal faz isso automaticamente. Entretanto, os arquivos com os dados de entrada do *workflow* devem ser carregados no Portal nesse momento (operação de *upload*). Para efetivamente executar os *workflows*, o componente de *Submissão* invoca o SciCumulus, que gerencia a execução dos *workflows*. Como o BioSciCumulus usa o banco de dados de proveniência do SciCumulus, a configuração da base de dados cria o esquema de dados público, que contém as tabelas e funções necessárias para a gerência dos dados de proveniência do *workflow* científico. Esse esquema segue o modelo de proveniência PROV-Df (Silva et al. 2016), que é compatível com o padrão W3C PROV.

Neste artigo considerou-se a configuração e submissão dos *workflows* no Portal utilizando-se o SciCumulus. Entretanto, cabe ressaltar que outros SWC, com capacidade de processamento paralelo, poderiam ser utilizados, desde que os parâmetros de entrada e saída dos componentes de Configuração e Consultas do BioSciCumulus fossem adaptados para contemplar os arquivos de configuração e o modelo de dados de proveniência do novo SWC.

Por último, o componente de *Processamento de Consultas* permite a análise de dados, uma vez que lista as bases de dados com *workflows* executados, permitindo que o usuário escolha uma base de dados a ser analisada. A Figura 2 mostra as tabelas da base *scc-sciphy* – base de dados obtida em nossos experimentos ao executar o *workflow* SciPhy usando o SciCumulus. Basta que o usuário escolha a tabela que deseja visualizar para que ela seja apresentada na interface *Web*.

Figura 2: Lista das tabelas do esquema *public* do *workflow scc-sciphy*

Na submissão de consultas do BioSciCumulus, o usuário pode submeter três tipos de consulta (uma vez que a base de proveniência do SciCumulus contém os dados que apoiam tais consultas): domínio, monitoramento e depuração. O usuário escolhe o tipo de consulta desejado e as consultas correspondentes já cadastradas são listadas. Em seguida, o usuário escolhe a consulta e faz a sua configuração, a partir de um formulário contendo os parâmetros necessários para a submissão da consulta. A Figura 3 apresenta uma consulta do tipo monitoramento. Após o término da configuração, a consulta é

submetida ao portal, para que ela seja executada sobre a base de dados cadastrada no portal, e o resultado é apresentado junto à descrição textual da consulta (Figura 4). No caso da utilização de outros SWC, um cliente é criado para se comunicar com o sistema de banco de dados do SWC, para que o portal seja usado com a mesma interface de submissão de consultas.

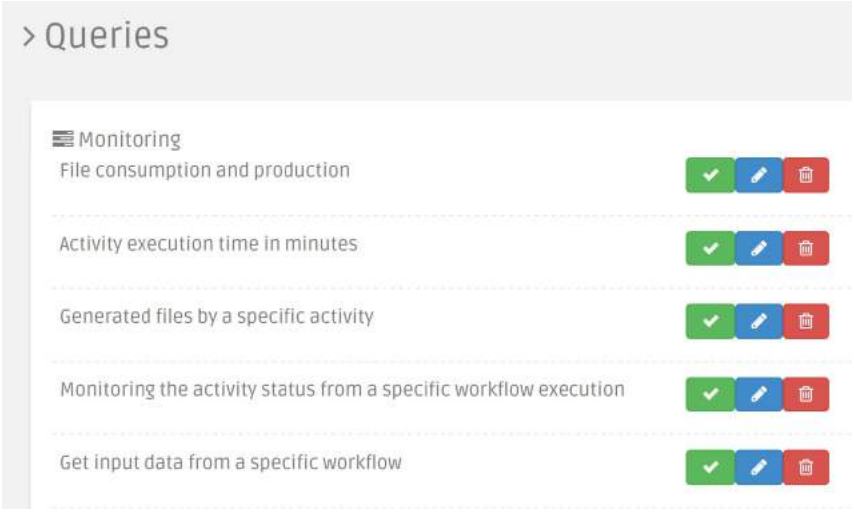

Figura 3: Consultas do tipo *Monitoramento*

Apesar de o BioSciCumulus já apresentar uma lista de consultas parametrizáveis cadastradas, as mesmas podem não atender a necessidade analítica do usuário. Nesse caso, é possível cadastrar e disponibilizar novas consultas, com o apoio de um especialista em computação. Para o cadastro de novas consultas, o Portal BioSciCumulus permite que os usuários informem uma descrição textual e a consulta na linguagem SQL para ser processada no banco de dados do portal. Uma vez cadastrada no portal, o usuário pode analisar a sua descrição textual e definir os parâmetros para a consulta de interesse. Como mencionado por Jagadish *et al.* (2007), o requisito facilidade de uso precisa ser tão importante quanto o desempenho. Eles observam que mesmo linguagens de consulta que permitem a elaboração de consultas complexas, ainda exigem um conhecimento detalhado do esquema da base de dados e esforços de programação para expressar as consultas. Considerando essas questões, observou-se que as principais consultas realizadas pelos usuários se concentravam em tabelas referentes às relações com dados de domínio (bioinformática) e às tarefas executadas (tabela *task*). Por isso, desenvolvemos visões que permitem ao usuário navegar no grafo de proveniência do experimento sem que o mesmo tenha que se preocupar com junções de várias tabelas do banco de dados. Essas visões podem ser utilizadas para gerar consultas parametrizáveis mais complexas que envolvem percorrer todo o grafo enquanto consideram também os dados produzidos.

> Monitoring the activity status from a specific workflow execution

tag	status	starttime	endtime
dataselection	FINISHED	2015-03-17 17:24:14.543-03	2015-03-17 17:24:16.57-03
mafft	FINISHED	2015-03-17 17:24:16.571-03	2015-03-17 17:24:19.409-03
readseq	FINISHED	2015-03-17 17:24:19.414-03	2015-03-17 17:24:22.489-03
modelgenerator	FINISHED	2015-03-17 17:24:22.491-03	2015-03-17 18:46:50.483-03
raxml1	FINISHED	2015-03-17 18:46:54.701-03	2015-03-17 19:30:36.358-03
raxml2	FINISHED	2015-03-17 18:46:50.484-03	2015-03-17 19:23:23.517-03
mergeRaxml	FINISHED	2015-03-17 19:30:36.364-03	2015-03-17 19:30:36.497-03
raxml3	FINISHED	2015-03-17 19:30:36.531-03	2015-03-17 19:30:43.135-03

Figura 4: Resultado da consulta de monitoramento

4. Conclusão

Os *workflows* científicos no cenário de Biologia Marinha computacional auxiliam os usuários em seus trabalhos diários de execução e análise de dados, desde o sequenciamento, anotação e análise de genes/genomas até filogenia e redes metabólicas. Apesar de representarem um avanço, os SWC existentes ainda oferecem apoio limitado no que tange à análise dos dados para não-especialistas em computação. Portais científicos possuem recursos mais voltados ao usuário não especialista, mas por serem muito genéricos, também possuem apoio limitado para consultas a dados em domínios específicos. Este artigo apresentou o Portal BioSciCumulus com uma instanciação no domínio da rede de pesquisas BiotecMar. A modelagem do domínio e execução via BioSciCumulus possibilitou a submissão de *workflows* científicos de uma forma mais simples em ambientes de PAD como *clusters* e nuvens e facilitou o monitoramento e análise da execução de *workflows*, com consultas sobre os dados de proveniência e de domínio. Uma vantagem do BioSciCumulus em comparação com as abordagens existentes é que ele não fixa as possíveis consultas, sendo um ambiente configurável por parte do usuário. Como trabalhos futuros, pretende-se proporcionar um mecanismo de geração automática de consultas em SQL, durante a fase de análise, de acordo com restrições especificadas pelo usuário de uma forma gráfica e sua incorporação a outros SWC, assim como a instalação e a configuração desta solução no portal da Rede Nacional de Bioinformática (http://bioinfo.lncc.br) alocado ao Sistema Nacional de Processamento de Alto Desempenho (https://www.lncc.br/sinapad) no LNCC.

Agradecimentos. Os autores gostariam de agradecer ao CNPq, Capes e FAPERJ pelo financiamento parcial deste trabalho.

Referências Bibliográficas

Abouelhoda, M., Issa, S., Ghanem, M., (2012), "Tavaxy: Integrating Taverna and Galaxy workflows with cloud computing support", *BMC Bioinformatics*, v. 13, p. 77.

Andrade, A. C., Fróes, A., Lopes, F. Á. C., Thompson, F. L., Krüger, R. H., Dinsdale, E., Bruce, T., (2017), "Diversity of Microbial Carbohydrate-Active enZYmes (CAZYmes) Associated with Freshwater and Soil Samples from Caatinga Biome", *Microbial Ecology* (Jan.)

Benza, S., Ocaña, K., Silva, V., Oliveira, D., Mattoso, M., (2015), "Modelling Data-intensive Metagenomics Experiments Using Scientific Workflows". In: *X-Meeting 2015 - 11th International Conference of the AB3C + Brazilian Symposium of Bioinformatics*, São Paulo.

Davidson, S. B., Freire, J., (2008), "Provenance and scientific workflows: challenges and opportunities". In: *ACM SIGMOD*, p. 1345–1350, Vancouver, Canada.

Gesing, S., Dooley, R., Pierce, M., Krüger, J., Grunzke, R., Herres-Pawlis, S., Hoffmann, A., (2017), "Gathering requirements for advancing simulations in HPC infrastructures via science gateways", *Future Generation Computer Systems* (Mar.)

Jagadish, H. V., Chapman, A., Elkiss, A., Jayapandian, M., Li, Y., Nandi, A., Yu, C., (2007), "Making database systems usable". , p. 13

Jain, A., Ong, S. P., Chen, W., Medasani, B., Qu, X., Kocher, M., Brafman, M., Petretto, G., Rignanese, G.-M., et al., (2015), "FireWorks: a dynamic workflow system designed for high-throughput applications", *CCPE*, v. 27, n. 17, p. 5037–5059.

Mattoso, M., Dias, J., Ocaña, K. A. C. S., Ogasawara, E., Costa, F., Horta, F., Silva, V., de Oliveira, D., (2015), "Dynamic steering of HPC scientific workflows: A survey", *FGCS*, v. 46 (May.), p. 100–113.

McLennan, M., Clark, S., Deelman, E., Rynge, M., Vahi, K., McKenna, F., Kearney, D., Song, C., (2015), "HUBzero and Pegasus: integrating scientific workflows into science gateways: HUBZERO AND PEGASUS", *Concurrency and Computation: Practice and Experience*, v. 27, n. 2 (Feb.), p. 328–343.

Nguyen, H. A., Abramson, D., Kipouros, T., Janke, A., Galloway, G., (2015), "WorkWays: interacting with scientific workflows", *CCPE*, v. 27, n. 16 (Nov.), p. 4377–4397.

Ocaña, K. A. C. S., Oliveira, D. de, Horta, F., Dias, J., Ogasawara, E., Mattoso, M., (2012), "Exploring Molecular Evolution Reconstruction Using a Parallel Cloud-based Scientific Workflow". In: *BSBBSB*, p. 179–191, Berlin, Heidelberg.

Ocaña, K. A. C. S., Oliveira, D., Dias, J., Ogasawara, E., Mattoso, M., (2011a), "Optimizing Phylogenetic Analysis Using SciHmm Cloud-based Scientific Workflow". In: *Proceedings of the 7th IEEE International Conference on e-Science (e-Science)IEEE e-Science 2011*, p. 190–197, Stockholm, Sweden.

Ocaña, K., Oliveira, D. de, Ogasawara, E., Dávila, A., Lima, A., Mattoso, M., (2011b), "SciPhy: A Cloud-Based Workflow for Phylogenetic Analysis of Drug Targets in Protozoan Genomes". In: *Advances in Bioinformatics and Computational Biology*, p. 66–70

Oliveira, D., Ogasawara, E., Baião, F., Mattoso, M., (2010), "SciCumulus: A Lightweight Cloud Middleware to Explore Many Task Computing Paradigm in Scientific Workflows". In: *International Conference on Cloud ComputingInternational Conference on Cloud Computing*, p. 378–385, Washington, DC, USA.

Reuillon, R., Leclaire, M., Rey-Coyrehourcq, S., (2013), "OpenMOLE, a workflow engine specifically tailored for the distributed exploration of simulation models", *Future Generation Computer Systems*, v. 29, n. 8 (Oct.), p. 1981–1990.

Silva, V., de Oliveira, D., Valduriez, P., Mattoso, M., (2016), "Analyzing related raw data files through dataflows", *CCPE*, v. 28, n. 8, p. 2528–2545.

Silva, V., Oliveira, D., Mattoso, M., (2014), "SciCumulus 2.0: Um Sistema de Gerência de Workflows Científicos para Nuvens Orientado a Fluxo de Dados". In: *Sessão de Demos do XXIX Simpósio Brasileiro de Banco de Dados*, Curitiba, Paraná.

Simon, C., Daniel, R., (2011), "Metagenomic Analyses: Past and Future Trends", *Applied and Environmental Microbiology*, v. 77, n. 4 (Feb.), p. 1153–1161.

Wozniak, J. M., Armstrong, T. G., Wilde, M., Katz, D. S., Lusk, E., Foster, I. T., (2013), "Swift/T: Large-Scale Application Composition via Distributed-Memory Dataflow Processing". In: *CCGrid*, p. 95–102