

Luciana da Silva Almendra Gomes

**Proveniência para Workflows de
Bioinformática**

Dissertação de Mestrado

Dissertação apresentada como requisito parcial para
obtenção do título de Mestre pelo Programa de Pós-
Graduação em Informática da PUC-Rio.

Orientador: Edward Hermann Haeusler

Rio de Janeiro

Abril de 2011



Luciana da Silva Almendra Gomes

**Proveniência para Workflows de
Bioinformática**

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre pelo Programa de Pós-graduação em Informática do Departamento de Informática do Centro Técnico e Científico da PUC-Rio. Aprovada pela Comissão Examinadora abaixo assinada.

Prof. Edward Hermann Haeusler

Orientador

Departamento de Informática – PUC-Rio

Prof. Sérgio Lifschitz

Departamento de Informática – PUC-Rio

Prof^a Marta Lima de Queirós Mattoso

UFRJ

Prof. Laurent Emmanuel Dardenne

LNCC

Prof. José Eugenio Leal

Coordenador Setorial do Centro

Técnico Científico – PUC-Rio

Rio de Janeiro, 27 de abril de 2011

Todos os direitos reservados. É proibida a reprodução total ou parcial do trabalho sem autorização do autor, do orientador e da universidade.

Luciana da Silva Almendra Gomes

Graduou-se em Informática na Pontifícia Universidade Católica do Rio de Janeiro (2008).

Ficha Catalográfica

Gomes, Luciana da Silva Almendra

Proveniência para Workflows de Bioinformática / Luciana da Silva Almendra Gomes ; orientador: Edward Hermann Haeusler. – 2011.
104 f. : il. ; 30 cm

Dissertação (mestrado)–Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Informática, 2011.
Inclui bibliografia

1. Informática – Teses. 2. Bioinformática. 3. Workflow. 4. Proveniência. I. Haeusler, Edward Hermann. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. III. Título.

CDD: 004

À minha avó Judith.
Aos meus pais, Lucia e Geraldo, e meu irmão, Vinicius.
Ao meu esposo, Adriano.
Às minhas filhas Beatriz e Alice.

Agradecimentos

A Deus, por tudo.

Aos meus pais, meus primeiros mestres, por todo o amor, dedicação e incentivo.

Ao meu irmão, um grande amigo, que sempre me ajudou em tudo.

Ao meu esposo, por todo o seu carinho, paciência, compreensão e colaboração nesses anos de estudo intenso.

Às nossas filhas, Beatriz e Alice, a primeira nascida no meio do mestrado, e a segunda um pouco depois da defesa. Beatriz e Alice chegaram para encher nossas vidas de alegria!

À minha família e meus amigos, que sempre me apoiaram em minhas decisões.

Ao meu orientador, Professor Sérgio Lifschitz, que desde a graduação direcionou os rumos da minha pesquisa, ajudando-me a distinguir entre um trabalho meramente técnico e uma pesquisa científica.

Aos professores que participaram da banca examinadora, por todos os excelentes comentários.

A Philippe Picouet e Sébastien Bigaret, pela cooperação com o nosso trabalho.

Aos amigos do LabBio, pelas inúmeras ajudas. À Márcia, por ter contribuído com tanta prontidão com a descrição de um dos estudos de caso. Aos alunos Waldecir e Diego, pela ajuda com a implementação do projeto.

Aos amigos Elaine e Hugo, pelas idéias e ajudas com o texto.

Aos professores e à secretaria de pós-graduação, por todo o serviço prestado com excelência e solicitude.

Ao CNPq, pelos auxílios concedidos, sem os quais não teria sido possível realizar esse trabalho.

Resumo

Gomes, Luciana da Silva Almendra; Haeusler, Edward Hermann. **Proveniência para Workflows de Bioinformática**. Rio de Janeiro, 2011. 104p. Dissertação de Mestrado – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Muitos experimentos científicos são elaborados como fluxos de tarefas computacionais, que podem ser implementados através do uso de linguagens de programação. Na área de bioinformática é muito comum o uso de *scripts ad-hoc* para construir fluxos de tarefas. Os Sistemas de Gerência de Workflow Científico (SGWC) surgiram como uma alternativa a estes *scripts*. Uma das funcionalidades desses sistemas que têm recebido bastante atenção pela comunidade científica é a captura automática de dados de proveniência. Estes permitem averiguar quais foram os recursos e parâmetros utilizados na geração dos resultados, dentre muitas outras informações indispensáveis para a validação e publicação de um experimento. Neste trabalho foram levantados alguns desafios na área de proveniência de dados em SGWCs, como por exemplo (i) a heterogeneidade de formas de representação dos dados nos diferentes sistemas, dificultando a compreensão e a interoperabilidade; (ii) o armazenamento de dados consumidos e produzidos e (iii) a reprodutibilidade de uma execução específica. Estes desafios motivaram a elaboração de um esquema conceitual de proveniência de dados para a representação de *workflows*. Foi implementada também uma extensão em um SGWC específico (BioSide) para incluir dados de proveniência e armazená-los utilizando o esquema conceitual proposto. Foram priorizados neste trabalho alguns requisitos comumente encontrados em *workflows* de Bioinformática.

Palavras-chave

bioinformática; workflow; proveniência.

Abstract

Gomes, Luciana da Silva Almendra; Haeusler, Edward Hermann (Advisor). **Provenance for Bioinformatics Workflows**. Rio de Janeiro, 2011. 104p. MSc. Dissertation – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Many scientific experiments are designed as computational workflows, which can be implemented using traditional programming languages. In the Bioinformatics domain *ad-hoc* scripts are often used to build workflows. Scientific Workflow Management Systems (SWMS) have emerged as an alternative to those scripts. One particular SWMS feature that has received much attention by the scientific community is the automatic capture of provenance data. These allow users to track which resources and parameters were used to obtain the results, among many other required information to validate and publish an experiment. In the present work we have elicited some data provenance challenges in the SWMS context, such as (i) the heterogeneity of data representation schemes that hinders the understanding and interoperability; (ii) the storage of consumed and produced data and (iii) the reproducibility of a specific execution. These challenges have motivated the proposal of a data provenance conceptual scheme for workflow representation. We have implemented an extension of a particular SWMS system (Bioside) to include provenance data and store them using the proposed conceptual scheme. We have focused on some requirements commonly found in bioinformatics' workflows.

Keywords

bioinformatics; workflow; provenance.

Sumário

1	Introdução	14
1.1.	Objetivos	15
1.2.	Organização do Texto	15
2	Preliminares	17
2.1.	Estado da Arte	17
2.1.1.	Workflows Científicos	17
2.1.2.	Sistemas de Gerência de Workflow Científico	18
2.1.2.1.	Características Gerais dos SGWC	19
2.1.2.1.1.	Composição	20
2.1.2.1.2.	Execução	22
2.1.2.1.3.	Análise	22
2.1.3.	Dados de Proveniência	22
2.1.3.1.	Classificações de Proveniência	23
2.1.3.2.	Open Provenance Model	24
2.1.3.3.	Sistemas e modelos de proveniência	26
2.1.3.3.1.	Vistrails	26
2.1.3.3.2.	Kepler	28
2.1.3.3.3.	Taverna	31
2.1.3.3.4.	BioSide	32
2.1.3.3.5.	REDUX	33
2.1.3.3.6.	e-BioFlow	34
2.2.	Motivação	35
2.2.1.	Workflows em Bioinformática	35
2.2.2.	Desafios de Proveniência	39
2.2.2.1.	Reprodutibilidade	39
2.2.2.2.	Gerência de Dados Consumidos e Produzidos	40
2.2.2.3.	Reuso de Dados	42
2.2.2.4.	Descrição e Gerência de Atividades	44
2.2.2.5.	Consultas sobre Grafos	47
2.2.2.6.	Facilidade de Interação	47

2.2.2.7. Interoperabilidade	48
2.2.2.8. Modelagem Conceitual	48
2.2.2.9. Atualização de Workflow	50
2.2.2.10. Modelo de Proveniência	50
2.2.3. Níveis de Reprodutibilidade	51
2.2.3.1. Reuso de Definição	51
2.2.3.2. Reprodutibilidade Estrita	52
2.3. Objetivos	54
3 Projeto de Proveniência	55
3.1. Esquema Conceitual de Proveniência	55
3.1.1. Atividades	57
3.1.2. Definição do Workflow	58
3.1.3. Execução do Workflow	60
3.1.4. Discussão sobre Parâmetros e Portas	63
3.2. Suporte aos Desafios	64
3.2.1. Reprodutibilidade	64
3.2.2. Gerência de Dados Consumidos e Produzidos	65
3.2.3. Reuso de Dados	67
3.2.4. Descrição e Gerência de Atividades	67
3.2.5. Modelo de Proveniência	68
3.3. Conclusão	69
4 Implementação	71
4.1. Modelagem	71
4.2. Estudos de Caso	75
4.2.1. Geração de Árvore Filogenética	75
4.2.2. MHOLline	77
4.2.2.1. Implementação Atual	77
4.2.2.2. Implementação no BioSide	78
4.3. Exemplos Práticos de Contribuições	80
4.3.1. Reprodutibilidade	84
4.3.1.1. Reuso de Definição	84
4.3.1.2. Reprodutibilidade Estrita	85
4.3.2. Gerência de Dados Consumidos e Produzidos	85
4.3.3. Reuso de Dados	86
4.3.4. Descrição e Gerência de Atividades	86

4.3.5. Modelo de Proveniência	87
4.4. Conclusão	89
5 Conclusão	91
5.1. Contribuições	92
5.2. Trabalhos Futuros	93
Referências	94
Anexo 1. Descrição das tabelas	101
Anexo 2. Exemplos de arquivos de especificação	103

Lista de figuras

Figura 1 – Interface Principal do Taverna	19
Figura 2 – Atividades e Passos	21
Figura 3 – Elementos e dependências que podem participar de um grafo OPM	26
Figura 4 – Geração de uma versão de workflow para cada alteração	27
Figura 5 – Consulta ao log do Vistrails para listar módulos executados	28
Figura 6 – Esquema de proveniência do Kepler	30
Figura 7 – Esquema de dados de proveniência do Taverna	32
Figura 8 – Níveis L0 e L2 dos esquemas do REDUX	34
Figura 9 – Esquema relacional do e-BioFlow, obtido em [Ooms, 2009]	35
Figura 10 – Workflow MHOLline	36
Figura 11 – Workflow de geração de árvores filogenéticas	37
Figura 12 – Workflows principal e de refinamento	43
Figura 13 – Definição das portas em módulo Vistrails de acesso ao BLAST	45
Figura 14 – Workflow que submete um arquivo local ao BLAST	45
Figura 15 – Erro na execução gerado pela inexistência do módulo Blast	46
Figura 16 – Esquema ER de proveniência de dados	56
Figura 17 – Descrição de Atividades	58
Figura 18 – Proveniência prospectiva	60
Figura 19 – Proveniência retrospectiva	62
Figura 20 – Ligações de proveniência	66
Figura 21 – Versionamento de atividades	68
Figura 22 – Diagrama de Classes	72
Figura 23 – Diagrama de Sequência	74
Figura 24 – Workflow para construção de árvore filogenética	75
Figura 25 – Arquitetura original do MHOLline	78
Figura 26 – MHOLline com chamada ao BioSide	79
Figura 27 – Workflow MHOLline implementado no BioSide	79
Figura 28 – Tabela Workflow	81
Figura 29 – Tabela Step	81
Figura 30 – Tabela Activity	81
Figura 31 – Tabela ExternalResource	82
Figura 32 – Tabela Parameter	82

Figura 33 – Tabela PortLink	82
Figura 34 – Tabela OPMGraph	82
Figura 35 – Tabela OPMPProcess	83
Figura 36 – Tabela OPMArtifact	83
Figura 37 – Tabela InputValue	83
Figura 38 – Tabela InputValue_Artifact	83
Figura 39 – Tabela Annotation	83
Figura 40 – Consultas 1, 2 e 3 em SQL	89
Figura 41 – Descrição das tabelas – parte 1	101
Figura 42 – Descrição das tabelas – parte 2	102
Figura 43 – Arquivo descritor do workflow de Geração de Árvore Filogenética	103
Figura 44 – Arquivo descritor da atividade Clustalw	104

Lista de tabelas

Tabela 1 – Objetivos do Projeto de Proveniência

54