1
Introdução

A Bioinformática é a área da computação que, entre outros objetivos, investiga o armazenamento e análise dos dados obtidos em experimentos científicos relacionados com a Biologia. Um pesquisador desta área utiliza recursos computacionais como programas e bancos de dados para processar as informações que foram geradas em bancada. Ao utilizar esses recursos, dizemos que o cientista está realizando um experimento *in-silico*, ou seja, um experimento realizado com o auxílio de ferramentas computacionais para testar uma hipótese, buscar padrões ou demonstrar fatos [Greenwood et al., 2003].

Muitos experimentos *in-silico* são compostos por uma sequência de programas encadeados, chamada comumente de pipeline ou workflow. Esse encadeamento possui a característica de data-flow, ou seja, é composto por fluxos de dados entre um programa e outro, tendo poucas ocorrências de fluxos de controle. Executar um workflow de tarefas sem o auxílio de uma ferramenta própria exige que o cientista execute cada programa separadamente fornecendo como entrada da tarefa seguinte a saída da anterior. Além de ser trabalhoso este processo possui vários inconvenientes. Por exemplo, se o usuário desejar testar seu workflow com uma nova entrada ou novos parâmetros terá que executar todo o procedimento novamente. Além disso, é um processo sujeito a erros, de difícil validação. Uma primeira solução para este problema foi o desenvolvimento *ad-hoc* de scripts para executar todo o pipeline desejado de uma só vez, o que automatiza o processo, facilita a reprodução do experimento e o torna menos sujeito a erros.

Em evolução aos scripts surgiram os sistemas de construção e gerência de workflows científicos, que chamaremos neste trabalho de SGWC. Estes sistemas permitem que o usuário construa seus fluxos de processos de forma intuitiva, sem a necessidade de saber programar um script. Em geral esses sistemas possuem uma interface gráfica, na qual o usuário pode adicionar facilmente os processos e concatená-los, gerando seu workflow. Além da construção, os SGWC permitem a execução do workflow gerado, o que faz deles uma ferramenta completa de construção, execução e visualização dos resultados do experimento.

Uma das funcionalidades de grande importância que esses sistemas podem oferecer é a captura automática de dados de proveniência, que são informações imprescindíveis para a validação do processo científico.

Nesta dissertação pretendemos levantar algumas questões de pesquisa relacionadas à proveniência de dados em SGWC, especialmente em relação a modelagem e armazenamento. Alguns requisitos elicitados foram escolhidos para o desenvolvimento de um projeto de proveniência de dados.

1.1. Objetivos

Neste trabalho será abordado o tema da proveniência de dados em SGWC, com ênfase nas necessidades de data-flows de Bioinformática. Os objetivos principais do trabalho são:

• O levantamento de alguns desafios atuais na área de proveniência em SGWC.

• A proposta de um modelo de proveniência motivada pelos desafios e por algumas necessidades de data-flows de Bioinformática.

• A implementação como prova de conceito utilizando dois estudos de caso da área da Bioinformática e o SGWC BioSide.

1.2. Organização do Texto

No capítulo PRELIMINARES são apresentados brevemente alguns conceitos sobre workflows, sistemas de gerência de workflow científico e proveniência de dados. Foram levantados alguns desafios atuais na área de proveniência para SGWC, dos quais escolhemos um subconjunto como motivação para o Projeto de Proveniência (Capítulo 3).

Em PROJETO DE PROVENIÊNCIA é descrito um projeto de proveniência de dados para SGWC. O projeto foi direcionado a data-flows, com ênfase no uso de programas de linha de comando. Apresentamos também neste capítulo como os desafios escolhidos foram abordados na modelagem proposta.

Em IMPLEMENTAÇÃO é descrita a especificação de uma extensão para o SGWC BioSide responsável por capturar os dados de proveniência e armazená-los em banco de dados relacional. Na seção Estudos de Caso são descritos dois

workflows de Bioinformática que foram implementados no sistema BioSide para demonstrar a utilização prática da implementação do projeto. Apresentamos também neste capítulo exemplos práticos de como os desafios escolhidos foram tratados na modelagem proposta.

Em CONCLUSÃO é feito um resumo da dissertação, são enumeradas as contribuições e listados alguns possíveis trabalhos futuros.