1 Introdução

1.1. Objetivos e Contribuição da Tese

Os projetos para estudo de genomas partem de uma fase de sequenciamento onde são gerados em laboratório dados brutos, ou seja, sequências de DNA sem significado biológico. As sequências de DNA possuem códigos responsáveis pela produção de proteínas e RNAs, enquanto que as proteínas participam de todos os fenômenos biológicos, como a replicação celular, produção de energia, defesa imunológica, contração muscular, atividade neurológica e reprodução. As sequências de DNA, RNA e proteínas são chamadas nesta tese de biossequências.

Como as biossequências possuem um papel fundamental em todos os organismos, espera-se que o seu entendimento leve a uma revolução em inúmeras áreas, como a medicina, biologia, agricultura, pecuária, entre outras.

O grande desafio dos pesquisadores consiste em analisar essas biossequências e obter informações biologicamente relevantes. Durante esta análise, os pesquisadores utilizam diversas ferramentas, programas de computador, e um grande volume de informações armazenadas em fontes de dados de Biologia Molecular. De fato, o crescente volume e a distribuição das fontes de dados e a implementação de novos processos em Bioinformática facilitaram enormemente a fase de análise. Porém, criaram uma demanda por ferramentas e sistemas semi-automáticos para lidar com tal volume e complexidade.

Esta tese aborda, em geral, a construção de sistemas que facilitem a fase de análise (de biossequências) e, em particular, o uso de workflows para compor processos de Bioinformática. Tais sistemas serão denominados na tese de *sistemas de gerência de análises em biossequências* (SGABios).

A tese apresenta inicialmente um levantamento de requisitos para SGABio's.

Em seguida, propõe um *framework* para um SGABio, organizado em módulos, definidos de acordo com suas responsabilidades, que atende aos requisitos levantados. O *framework* decompõe um SGABio em dois sub-sistemas: um *sistema de gerência de workflows de Bioinformática*, que auxilia os pesquisadores na definição, validação, otimização e execução de workflows necessários para se realizar as análises; e um *sistema de gerência de dados em Bioinformática*, que trata do armazenamento e da manipulação dos dados envolvidos nestas análises.

O *framework* inclui ainda um gerente de ontologias, armazenando ontologias para Bioinformática, que deverão ser utilizadas (nas instanciações do *framework*) para auxiliar o pesquisador desde a definição até a execução do workflow de forma otimizada, sem exigir que o pesquisador tenha conhecimentos avançados em Bioinformática ou em técnicas de otimização.

A tese apresenta uma ontologia particular, derivada de um estudo cuidadoso, que modela os principais processos, recursos, dados, projetos e ambientes de trabalho comumente envolvidos em análises de biossequências.

A principal contribuição da tese consiste em instanciações do *framework* para três tipos de ambiente de trabalho comumente encontrados e sugestivamente chamados de ambiente pessoal, ambiente de laboratório e ambiente de comunidade. Para cada um destes ambientes, a tese discute em detalhe os aspectos particulares da execução e otimização de workflows. O foco principal da discussão está nos algoritmos de gerência de execução, que incluem heurísticas que visam a otimização do workflow. As heurísticas foram definidas de acordo com as características particulares dos programas de Bioinformática e de cada ambiente em que o SGWBio pode ser implementado. Conforme será justificado na tese, existirão otimizações que poderão ser feitas no workflow *a priori* (antes de sua execução) e outras que só poderão ser definidas *a posteriori* (durante a execução).

Por fim, a tese descreve um protótipo implementado por instanciação dos pontos de flexibilização do *framework*. Apesar do protótipo ter sido desenvolvido para o ambiente pessoal, ele serve de modelo para implementações em outros ambientes de trabalho, já que muitas funcionalidades dos módulos do *framework* são comuns a todos os ambientes.

1.2. Organização da Tese

A tese está organizada em 8 capítulos.

O Capítulo 2 contém uma discussão sucinta dos contextos biológico e computacional, necessária para o entendimento e motivação deste trabalho. Esta discussão dá uma visão geral dos assuntos tratados em Bioinformática relevantes a esta tese, como os programas de análise, os bancos de dados, os sistemas de anotação, a integração dos dados e aplicativos, as ontologias e os workflows. O capítulo mostra a importância de um SGABio, especificando os requisitos que ele deve atender e termina com um levantamento dos principais trabalhos relacionados.

O Capítulo 3 apresenta a ontologia de processos de Bioinformática que direciona o SGABio, permitindo que o sistema seja capaz de auxiliar os pesquisadores a definir, redefinir, validar, otimizar e executar o workflow de forma adequada.

O Capítulo 4 define a linguagem utilizada pelo sistema de gerência de workflows para descrever workflows de Bioinformática. O documento de especificação de um workflow de Bioinformática refletirá o workflow definido por um pesquisador e conterá informações adicionais, como instâncias de processos extras, que permitirão que o workflow seja executado de forma coerente e otimizada.

O Capítulo 5 propõe um *framework* para um SGABio, enfatizando o sistema de gerência de workflows. O capítulo descreve os módulos do *framework*, seus pontos de flexibilização e suas responsabilidades, de acordo com os requisitos levantados no Capítulo 2. Este capítulo indica a importância da ontologia definida no Capítulo 3 e a utilização do documento de especificação do workflow, de acordo com a linguagem definida no Capítulo 4. Como o *framework* pode ser instanciado em diferentes ambientes de trabalho dos pesquisadores, este capítulo mostra as suas características gerais, ou seja, pertinentes a todos os ambientes.

O Capítulo 6 define as arquiteturas dos ambientes de trabalho mais comuns dos pesquisadores e discute em considerável detalhe instanciações do sistema de gerência de workflows para estes ambientes.

O Capítulo 7 apresenta o protótipo de um sistema de gerência de workflows para o ambiente pessoal, construído por instanciação dos *hot spots* do *framework* definido no Capítulo 5.

Por fim, o Capítulo 8 resume o trabalho realizado, compara-o com os trabalhos relacionados mais relevantes, apresenta as contribuições e sugere trabalhos futuros.