Pontifícia Universidade Católica
do Rio de Janeiro

Melissa Lemos

Workflow para Bioinformática

Tese de Doutorado

Tese apresentada ao Programa de Pós-graduação em Informática da PUC-Rio como requisito parcial para obtenção do título de Doutor em Informática.

Orientador: Prof. Marco Antonio Casanova

Co-Orientador: Prof. Antônio Basílio de Miranda

Rio de Janeiro, Setembro de 2004

PONTIFÍCIA UNIVERSIDADE CATÓLICA
DO RIO DE JANEIRO

Melissa Lemos

Workflow para Bioinformática

Tese apresentada como requisito parcial para obtenção do título de Doutor pelo Programa de Pós-Graduação em Informática da PUC-Rio. Aprovada pela Comissão Examinadora abaixo assinada.

Prof. Marco Antonio Casanova
(Orientador)
Departamento de Informática – PUC-Rio

Prof. Antônio Basílio de Miranda
(Co-Orientador)
Departamento de Bioquímica – Fiocruz

Profa. Marta L. Queirós Mattoso
Departamento de Informática – UFRJ

Profa. Ana Maria de Carvalho Moura
Departamento de Informática – IME-RJ

Prof. Paulo Mascarello Bisch
Departamento de Bioquímica – UFRJ

Prof. Luiz Fernando Bessa Seibel
Departamento de Informática – PUC-Rio

Prof. Marcus V. S. Poggi de Aragão
Departamento de Informática – PUC-Rio

Prof. Rubens Nascimento Melo
Departamento de Informática – PUC-Rio

Prof. José Eugênio Leal
Coordenador Setorial do Centro Técnico Científico – PUC-Rio

Rio de Janeiro, 13 de Setembro de 2004

Melissa Lemos

Graduou-se em Engenharia de Computação na PUC-Rio (Pontifícia Universidade Católica do Rio de Janeiro) em 1998. Obteve o título de Mestre em Informática na PUC-Rio em 2000, tendo a Dissertação de Mestrado ênfase em Banco de Dados e Bioinformática. Apresentou trabalhos sobre Bioinformática em congressos. É uma das pesquisadoras responsáveis pelo BioNotes, um sistema de anotação de sequências que está sendo utilizado no Estado do Rio de Janeiro para a pesquisa de genomas de organismos.

Ficha Catalográfica

Lemos, Melissa

Workflow para bioinformática / Melissa Lemos ; orientador: Marco Antonio Casanova ; co-orientador: Antônio Basílio de Miranda. - Rio de Janeiro : PUC, Departamento de Informática, 2004.

239 f. : il. ; 30 cm

Tese (doutorado) – Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Informática.

Inclui referências bibliográficas.

CDD: 004

Ao meu marido, Marcio

Agradecimentos

A Deus, por iluminar-me e abençoar-me sempre.

Ao meu querido marido, Marcio Barroso Cavaliére, a quem tanto amo, pelo seu amor, carinho, torcida, companheirismo durante este longo caminho, paciência e compreensão nas inúmeras ocasiões em que eu não pude dar-lhe a merecida atenção e por tornar meus dias mais felizes.

Ao meu pai, Mauro César de Lemos, por todo amor e carinho a mim dedicado e por me ensinar a importância do estudo e do conhecimento.

A minha mãe, Alaide das Graças Lemos, pelo amor, carinho, presença em todos os momentos da minha vida e apoio incondicional em todos os meus sonhos.

Ao meu irmão, Frederico Lemos, amigo que sempre está ao meu lado, torce por mim, acompanha e incentiva meus estudos.

Ao meu orientador, Marco Antonio Casanova, que tem minha eterna gratidão e admiração, pela excelente orientação, por sua dedicação e disponibilidade incansáveis, por suas idéias brilhantes, pela credibilidade depositada em mim, pela serenidade de nossas conversas que não me deixavam desanimar, por seus ensinamentos e pela parceria imprescindível para a realização desta tese.

Ao meu co-orientador, Antônio Basílio de Miranda, pela disponibilidade, atenção e paciência infinita em esclarecer tantas dúvidas em Biologia, contribuindo para a solidificação de uma base necessária para a pesquisa realizada nesta tese.

Ao Luiz Fernando Bessa Seibel, amigo que tem meu reconhecimento e gratidão, pela amizade e otimismo que me incentivaram e fizeram acreditar no valor do meu trabalho, pelas inúmeras sugestões, contribuições e conversas que me ajudaram a ter idéias, a criar e a melhorar a tese, e pelas oportunidades de trabalho, que me permitiram crescer profissionalmente e realizar esta tese.

Ao meu amigo José Antônio Fernandes de Macedo, a quem tenho como exemplo de amizade e sabedoria, agradeço pelo bom humor contagiante, pela disponibilidade, atenção e participação, e por sempre procurar me ajudar e enriquecer este trabalho com suas valiosas críticas e idéias.

Aos amigos Michelle Santos Sá, Flávio Freitas, Rogério Costa e Valéria Bastos, por sempre me passarem alegria, otimismo e pelas sugestões e críticas que tornaram o trabalho melhor.

Ao prof. Warren Gish, que sempre respondeu minhas perguntas por e-mails me esclarecendo dúvidas sobre o BLAST. Sem estes esclarecimentos, não teria sido possível propor várias idéias apresentadas nesta tese.

A todos os profissionais que estiveram envolvidos na especificação e no desenvolvimento do BioNotes, com quem eu tive a felicidade de conviver durante estes anos. Agradeço aos pesquisadores do Consórcio Riogene, em especial ao Paulo Ferreira Cavalcanti, Orlando Martins, Marcelo Bertalan, Shaila Rossle, Paulo Bisch e Ana Coelho. Agradeço aos profissionais em Informática que participaram, sempre com muito entusiasmo e empenho, no desenvolvimento do BioNotes. Meu reconhecimento aos colegas Fernando Mano, Vitor Cruz, Carlos Eduardo Vieira, Roberto Cavalcante, Dan Eisenberg e Andrea Cynthia dos Santos.

Aos professores Paulo Bisch, Marta Mattoso, Ana Maria Moura, Luiz Fernando Seibel, Marcus Poggi de Aragão e Rubens Melo pela participação na banca de avaliação.

A minha família e amigos pelo estímulo, apoio e orações.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)
pelo auxílio que me foi concedido.

Resumo

Lemos, Melissa; Casanova, Marco Antonio (Orientador) ; Miranda, Antonio Basílio de (Co-orientador). **Workflow para Bioinformática**. Rio de Janeiro, 2004, 239p. Tese de Doutorado – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

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. Porém, o grande desafio destes projetos consiste em analisar essas biossequências, e obter informações biologicamente relevantes. Durante a fase de análise, os pesquisadores usam diversas ferramentas, programas de computador, e um grande volume de informações armazenadas em fontes de dados de Biologia Molecular. 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. Neste cenário, esta tese aborda o uso de workflows para compor processos de Bioinformática, facilitando a fase de análise. Inicialmente apresenta uma ontologia modelando processos e dados comumente utilizados em Bioinformática. Esta ontologia foi derivada de um estudo cuidadoso, resumido na tese, das principais tarefas feitas pelos pesquisadores em Bioinformática. Em seguida, a tese propõe um *framework* para um sistema de gerência de análises em biossequências, composto por dois sub-sistemas. O primeiro é 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. O segundo é um sistema de gerência de dados em Bioinformática, que trata do armazenamento e da manipulação dos

dados envolvidos nestas análises. O *framework* inclui um gerente de ontologias, armazenando ontologias para Bioinformática, nos moldes da apresentada anteriormente. Por fim, a tese descreve 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.

Palavras-chave

Bioinformática; Banco de Dados; Workflow; Ontologia; Framework de Software.

Abstract

Lemos, Melissa; Casanova, Marco Antonio (Advisor); Miranda, Antonio Basílio de (Co-Advisor). **Workflow for Bioinformatics**. Rio de Janeiro, 2004, 239p. PhD. Thesis – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Genome projects usually start with a sequencing phase, where experimental data, usually DNA sequences, is generated, without any biological interpretation. DNA sequences have codes which are responsible for the production of protein and RNA sequences, while protein sequences participate in all biological phenomena, such as cell replication, energy production, immunological defense, muscular contraction, neurological activity and reproduction. DNA, RNA and protein sequences are called biosequences in this thesis. The fundamental challenge researchers face lies exactly in analyzing these sequences to derive information that is biologically relevant. During the analysis phase, researchers use a variety of analysis programs and access large data sources holding Molecular Biology data. The growing number of Bioinformatics data sources and analysis programs indeed enormously facilitated the analysis phase. However, it creates a demand for systems that facilitate using such computational resources. Given this scenario, this thesis addresses the use of workflows to compose Bioinformatics analysis programs that access data sources, thereby facilitating the analysis phase. An ontology modeling the analysis program and data sources commonly used in Bioinformatics is first described. This ontology is derived from a careful study, also summarized in the thesis, of the computational resources researchers in Bioinformatics presently use. A framework for biosequence analysis management systems is next described. The system is divided into two major components. The first component is a Bioinformatics workflow management system that helps researchers define, validate, optimize and run workflows combining Bioinformatics analysis programs. The second component is a Bioinformatics data management system that helps researchers manage large volumes of Bioinformatics data. The framework includes an ontology manager that stores Bioinformatics ontologies, such as that previously described. Lastly,

instantiations for the Bioinformatics workflow management system framework are described. The instantiations cover three types of working environments commonly found and suggestively called personal environment, laboratory environment and community environment. For each of these instantiations, aspects related to workflow optimization and execution are carefully discussed.

Keywords

Bioinformatics; Databases; Software Frameworks; Workflow; Ontology.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Quadros