PONTIFÍCIA UNIVERSIDADE CATÓLICA
DO RIO DE JANEIRO

Maíra Ferreira de Noronha

Controle da Execução e Disponibilização de Dados para Aplicativos sobre Seqüências Biológicas: o Caso BLAST

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: Sérgio Lifschitz

Rio de Janeiro,

setembro de 2006

Pontifícia Universidade Católica
do Rio de Janeiro

Maíra Ferreira de Noronha

Controle da Execução e Disponibilização de Dados para Aplicativos sobre Seqüências Biológicas: o Caso BLAST

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

Prof. Sérgio Lifschitz
Orientador
Departamento de Informática - PUC-Rio

Prof. Marcus Vinicius Soledade Poggi de Aragão
Departamento de Informática - PUC-Rio

Prof. Melissa Lemos Cavalieri
Departamento de Informática - PUC-Rio

Prof. Marta Lima de Queiros Mattoso
COPPE - UFRJ

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

Rio de Janeiro, 20 de setembro de 2006

Maíra Ferreira de Noronha

Graduou-se em Engenharia de Computação na Pontifícia Universidade Católica do Rio de Janeiro em 2003. Participou do Programa de Graduação Sanduíche da CAPES de 2001 a 2002, tendo realizado um ano do curso de Informática e Matemática Aplicada no ENSEEIHT, em Toulouse, França. Trabalha atualmente como Analista de Sistemas na Petrobrás.

Ficha Catalográfica

Noronha, Maíra Ferreira de

Controle da execução e disponibilização de dados para aplicativos sobre seqüências biológicas: o caso BLAST / Maíra Ferreira de Noronha ; orientador: Sérgio Lifschitz. – 2006.
83 f. : il. ; 30 cm

Dissertação (Mestrado em Informática)–Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2006.
Inclui bibliografia

1. Informática – Teses. 2. Banco de dados. 3. Gerência de Buffer. 4. Gerência de memória. 5. Bioinformática. 6. Biologia computacional. 7. Driver. 8. Escalonador de processos. I. Lifschitz, Sérgio. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. III. Título.

CDD: 004

Agradecimentos

Ao meu orientador, Sérgio Lifschitz, pelos inúmeros conselhos, pela confiança e pela paciência demonstrada ao longo dos anos.

À minha família, da qual recebo sempre muito carinho e apoio.

Ao meu namorado Marcelo por seu carinho, incentivo e apoio incondicional em todos os momentos.

Aos meus amigos, que tornam os dias sempre alegres e especiais.

Aos meus colegas da PUC-Rio, que foram muito receptivos e tornaram agradáveis os momentos na universidade.

Ao Daniel pela ajuda em diversos momentos durante o desenvolvimento da minha dissertação.

Ao CNPQ e à PUC-Rio pelos auxílios concedidos, sem os quais este trabalho não poderia ter sido realizado.

Resumo

Noronha, Maíra Ferreira de; Lifschitz, Sérgio. **Controle da Execução e Disponibilização de Dados para Aplicativos sobre Seqüências Biológicas: o Caso BLAST.** Rio de Janeiro, 2007. 83p. Dissertação de Mestrado - Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Este trabalho consiste na criação de uma ferramenta provedora de dados para o BLAST, denominada BioProvider. Esta é usada para prover dados realizando um gerenciamento de *buffer* eficiente para o BLAST, controlando também o escalonamento dos processos do mesmo. A comunicação entre o BioProvider e os processos do BLAST, assim como o controle de concorrência e bloqueios, é feita por meio de um *driver*, que substitui as chamadas a funções de leitura e escrita de arquivos do banco de dados. Deste modo, o código do BLAST não precisa ser modificado para ser realizar a comunicação com o BioProvider e este pode ser usado para diferentes versões do BLAST. O desenvolvimento do BioProvider é a primeira etapa para a criação de uma solução aplicável também a outras ferramentas de Bioinformática. Por ser transparente aos programas, a ferramenta desenvolvida é facilmente extensível, podendo ser futuramente modificada para prover dados para outros aplicativos, usar outras estratégias de gerência de *buffer* ou prover dados armazenados em formatos diferentes dos lidos por processos clientes, convertendo-os em tempo de execução. O BioProvider foi testado com a versão recente do NCBI BLAST, obtendo consideráveis melhoras de desempenho, e seu funcionamento foi verificado também com a versão do WU-BLAST com código aberto. Foram realizadas análises de variações no algoritmo de gerenciamento de *buffer* e dos fatores que influenciam o desempenho dos processos BLAST.

Palavras-chave

Bancos de Dados;Gerência de *Buffer*;Gerência de Memória;NCBI BLAST;WU-BLAST;Bioinformática;Biologia Computacional;Driver;Escalonador de Processos.

Abstract

Noronha, Maíra Ferreira de; Lifschitz, Sérgio. **Controle da Execução e Disponibilização de Dados para Aplicativos sobre Seqüências Biológicas: o Caso BLAST.** Rio de Janeiro, 2007. 83p. Dissertação de Mestrado - Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

This work consists on the creation of a tool named BioProvider to provide data to BLAST. The tool provides the data using buffer management techniques that are efficient for BLAST and controls process scheduling. The communication between BioProvider and the BLAST processes, as well as the concurrency and blocking control, is done through a device driver that substitutes the read and write function calls to the database files. By this means, the application code can remain unchanged and BioProvider can be used with different versions of BLAST. The development of BioProvider was the first stage to the creation of a solution that can be applied as well to other Bioinformatics tools. Due to its transparency in the view of other applications, BioProvider can be easily extended in the future to provide data to other applications, to use other buffer management techniques or to provide data stored in different formats of those read by the client processes, converting the data in runtime. BioProvider has been tested with the most recent version of NCBI BLAST and considerable improvement has been verified. The tool has been shown to work as well with the open source version of WU-BLAST. Some variations on the buffer management algorithm were studied, as well as the different factors that influence the performance of BLAST processes.

Keywords

Databases;Buffer Management;Memory Management;NCBI BLAST;WU-BLAST;Bioinformatics;Computational Biology;Driver;Data Scheduler.

Sumário

Lista de figuras

Lista de tabelas