



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



Máira 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

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

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 Sequências 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 Sequências 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

1	Introdução	13
2	Contexto e Motivações	16
2.1.	A Família BLAST	17
2.1.1.	Principais Características	18
2.1.2.	Subprogramas	19
2.2.	Formatos dos Bancos de Dados	20
2.2.1.	O Banco de Dados do NCBI BLAST	21
2.2.2.	O Banco de Dados do WU-BLAST	23
2.3.	O Acesso do BLAST ao Banco de Dados	24
2.3.1.	NCBI BLAST com 1 Seqüência de Entrada	24
2.3.2.	NCBI BLAST com 5 Seqüências de Entrada	26
2.3.3.	WU-BLAST com 1 Seqüência de Entrada	27
2.4.	Trabalhos de Melhoria do BLAST	27
2.4.1.	Gerenciamento de <i>Buffer</i> para o BLAST	29
2.4.2.	Gerenciamento de <i>Buffer</i> Não-Intrusivo	31
2.5.	Considerações Finais	32
3	O BioProvider	34
3.1.	A Estratégia Utilizada	35
3.1.1.	Gerenciamento de Buffer	36
3.1.2.	Tratamento do Banco de dados	37
3.1.3.	Melhorias no Desempenho	40

3.2. Implementação	43
3.3. Considerações Finais	47
4 Resultados Obtidos	48
4.1. Variação do Tamanho da Memória	50
4.2. Variação do Tamanho do Anel em Memória	52
4.3. Variação das Seqüências de Entrada	55
4.4. Variação do Número de Blocos	57
4.5. Variação da Estratégia de Atendimento dos Processos	60
4.6. Considerações Finais	62
5 Conclusões	64
5.1. Trabalhos Futuros	65
Referências Bibliográficas	67
A Funcionamento do BLAST	70
B Gerência de Memória em Bancos de Dados	73
C <i>Drivers</i> de Dispositivos para o Linux	75
C.1. Números Maior e Menor	76
C.2. Operações em um Arquivo Especial	77
C.3. Alocação de Memória	79
C.4. Execução de Tarefas	79
C.4.1. Execução em resposta a uma solicitação de um processo	80
C.4.2. Execução em resposta a interrupções	80
C.4.3. Execução espontânea	81
C.5. Controle de Concorrência	81

C.6. Bloqueio de Processos	82
C.7. Estrutura Básica de um Módulo do Linux	82

Lista de figuras

Figura 1: Leitura realizada pelo NCBI BLAST com 1 seqüência de entrada	25
Figura 2: Leitura realizada pelo NCBI BLAST com 5 seqüências de entrada	26
Figura 3: Leitura realizada pelo WU-BLAST com 1 seqüência de entrada	27
Figura 4: Acesso dos processos BLAST às páginas do anel em memória	29
Figura 6: Erro dos ponteiros do arquivo de índices.	38
Figura 7: Pré-formatação do banco de dados	39
Figura 8: Gráfico dos tempos médios, teste 1	51
Figura 9: Gráfico dos tempos totais, teste 1	51
Figura 10: Gráfico dos números de <i>page faults</i> , teste 1	51
Figura 11: Gráfico dos números de mudanças de contexto, teste 1	52
Figura 12: Gráfico dos tempos médios, teste 2	53
Figura 13: Gráfico dos tempos totais, teste 2	53
Figura 14: Gráfico dos números de <i>page faults</i> , teste 2	54
Figura 15: Gráfico dos números de mudanças de contexto, teste 2	54
Figura 16: Gráfico dos tempos médios, teste 3	55
Figura 17: Gráfico dos tempos totais, teste 3	56
Figura 18: Gráfico dos números de <i>page faults</i> , teste 3	56
Figura 19: Gráfico dos números de mudanças de contexto, teste 3	56
Figura 20: Gráfico dos tempos médios, teste 4	58
Figura 21: Gráfico dos tempos totais, teste 4	58
Figura 22: Gráfico dos números de <i>page faults</i> , teste 4	59
Figura 23: Gráfico dos números de mudanças de contexto, teste 4	59

Figura 24: Gráfico dos tempos médios, teste 5	61
Figura 25: Gráfico dos tempos totais, teste 5	61
Figura 26: Gráfico dos números de <i>page faults</i> , teste 5	61
Figura 27: Gráfico dos números de mudanças de contexto, teste 5	62

Lista de tabelas

Tabela 1: Códigos dos caracteres de aminoácidos do NCBI 2.0	22
Tabela 2: Códigos dos caracteres de aminoácidos do WU-BLAST 1.4	23
Tabela 3: Variáveis usadas nos testes e seus valores	49
Tabela 4: Valores das variáveis no teste 1	50
Tabela 5: Valores das variáveis no teste 2	53
Tabela 6: Valores das variáveis no teste 3	55
Tabela 7: Valores das variáveis no teste 4	58
Tabela 8: Valores das variáveis no teste 5	60