PONTIFÍCIA UNIVERSIDADE CATÓLICA
DO RIO DE JANEIRO

Janaina Oleinik Moura Rosa

Um Estudo de Compactação de Dados para Biosseqüências

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

Janaina Oleinik Moura Rosa

Um Estudo de Compactação de Dados para Biosseqüências

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.

Sérgio Lifschitz
Orientador
PUC-Rio

Fernanda Araújo Baião
UNIRIO

Luiz Fernando Bessa Seibel
PUC-Rio

Rubens Nascimento Melo
PUC-Rio

Prof. José Eugenio Leal
Coordenador(a) Setorial do Centro Técnico Científico - PUC-Rio

Rio de Janeiro, 6 de setembro de 2006

Janaina Oleinik Moura Rosa

Graduou-se em Engenharia Informática na Universidade Católica "Nstra. Sra de la Asunción", PY, em 1998. Atuou em empresas como Analista de Sistemas e Administrador de dados. Lecionou no curso de Administração e Tuning de Banco de Dados do CCE PUC-Rio.

Ficha Catalográfica

Rosa, Janaina Oleinik Moura

Um estudo de compactação de dados para biossequências / Janaina Oleinik Moura Rosa ; orientador: Sérgio Lifschitz. – Rio de Janeiro : PUC, Departamento de Informática, 2006.

135 f. : il. ; 30 cm

Dissertação (mestrado) – Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Informática.

Inclui bibliografia.

1. Informática – Teses. 2. BLAST. 3. Compactação. 4. Bioinformática. I. Lifschitz, Sérgio. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. III. Título.

CDD: 004

Aos meus pais pelo apoio incondicional. Ao meu marido, José Antônio, por incentivo e carinho constantes. E ao meu pequeno Antônio que chegou para dar mais alegria e motivação.

Agradecimentos

Ao professor Sérgio Lifschitz pela orientação e empenho na leitura e revisão do texto.

Ao professor Eduardo Laber pela colaboração com materiais e dicas sobre algoritmos de compressão de dados.

Ao meu marido José Antônio pelas inúmeras contribuições ao trabalho. Suas revisões e orientações foram fundamentais. Agradecimentos pela companhia carinhosa, mesmo que virtual em alguns momentos, não importando nem mesmo o fuso-horário. E principalmente por acreditar que seria possível.

Ao amigo Eduardo Morelli pelo companheirismo e amizade desde as primeiras disciplinas até a entrega deste documento, e principalmente pelas palavras otimistas nos momentos difíceis.

À amiga Michelle, pelo seu ponto de vista sempre tão prático e espirituoso.

Aos amigos Simone, Ciro e Maíra pelos momentos de descontração.

À minha família, meus pais, Rui e Leila, e irmãos, Ruizinho, Rafaela e Gabriela pelo amor e apoio incondicional em todos os momentos.

Ao meu filho Antônio que trouxe motivação e alegria extras na reta final.

Resumo

Oleinik Moura Rosa, Janaina; Lifschitz, Sérgio. **Um Estudo de Compactação de Dados para Biosseqüências.** Rio de Janeiro, 2006. 135p. Dissertação de Mestrado - Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

A família de algoritmos BLAST é a mais utilizada pelos biólogos para a busca de similaridade entre biosseqüências, e por esta razão, melhoras nestes algoritmos, em suas estruturas de dados ou em seus métodos de acesso à memória secundária são muito importantes para o avanço das descobertas biológicas. Nesta dissertação, foi estudada detalhadamente uma versão do programa BLAST, analisando as suas estruturas de dados e os algoritmos que as manipulam. Além disso, foram realizadas medições de desempenho com o intuito de identificar os possíveis gargalos de processamento dentro das fases de execução do BLAST. A partir das informações obtidas, técnicas de compactação de dados foram utilizadas como uma estratégia para redução de acesso à memória secundária com o objetivo de melhorar o desempenho para a execução do BLAST. Finalmente, foi gerada uma versão modificada do BLAST no ambiente Windows, na qual foi alterado diretamente o código do programa. Os resultados obtidos foram comparados com os resultados obtidos na execução do algoritmo original.

Palavras-chave

BLAST; compactação; Bioinformática

Abstract

Oleinik Moura Rosa, Janaina; Lifschitz, Sérgio. A Study of Biosequence Data Compression**.** Rio de Janeiro, 2006. 135p. Master Thesis - Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

The BLAST is the sequence comparison strategy mostly used in computational biology. Therefore, research on data structures, secondary memory access methods and on the algorithm itself, could bring important optimizations and consequently contributions to the area. In this work, we study a NCBI BLAST version by analyzing its data structures and algorithms for data manipulating. In addition, we collect performance data for identifying processing bottleneck in all the BLAST execution phases. Based on this analysis, data compress techniques were applied as a strategy for reducing number of secondary memory access operations. Finally, a modified version of BLAST was implemented in the Microsoft Windows environment, where the program was directly altered. Finally, an analysis was made over using the results of execution of original BLAST against modified BLAST.

Keywords

BLAST; data compression; Bioinformatics

Sumário

Lista de Tabelas

Lista de Figuras