



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



Janaina Oleinik Moura Rosa

Um Estudo de Compactação de Dados para Biossequências

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

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

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

| | | |
|-------|---|----|
| 1 | Introdução | 13 |
| 1.1 | Bioinformática e Banco de Dados | 13 |
| 1.2 | Objetivos da Dissertação | 14 |
| 1.3 | Estrutura da Dissertação | 15 |
| 2 | Preliminares | 17 |
| 2.1 | Contexto Biológico | 17 |
| 2.2 | Operações sobre Biosseqüências | 18 |
| 2.3 | Persistência de Biosseqüências | 20 |
| 2.4 | Problemas com o armazenamento de biosseqüências | 21 |
| 2.5 | Trabalhos relacionados | 23 |
| 2.6 | Conclusão | 25 |
| 3 | Análise do programa BlastP | 27 |
| 3.1 | Descrição do Funcionamento do BLAST | 27 |
| 3.2 | Descrição da implementação do NCBI-BLAST | 30 |
| 3.3 | Análise do desempenho | 34 |
| 3.3.1 | Metodologia de Análise | 34 |
| 3.3.2 | Análise | 37 |
| 3.4 | Conclusão | 43 |
| 4 | Compactação de dados | 45 |
| 4.1 | Compactação | 45 |
| 4.1.1 | Classificações | 47 |
| 4.1.2 | Algoritmos de compressão reversível | 49 |
| 4.2 | Compactação em SGBDs | 51 |
| 4.3 | Compactação de biosseqüências | 55 |
| 4.4 | Considerações Finais | 57 |
| 4.5 | Conclusão | 58 |
| 5 | Uma proposta de compactação para o BlastP | 60 |
| 5.1 | A solução proposta | 61 |

| | | |
|-------|---|-----|
| 5.2 | Detalhamento da Proposta de Solução | 63 |
| 5.3 | Conclusão | 70 |
| 6 | Resultados experimentais | 71 |
| 6.1 | Metodologia de análise de resultados | 71 |
| 6.2 | Análise de resultados | 76 |
| 6.2.1 | Resultados e Análises do Cenário 1- Testes de T1 a T5 | 77 |
| 6.2.2 | Resultados e Análises do Cenário 2 – Testes de T6 a T10 | 82 |
| 6.3 | Conclusão | 86 |
| 7 | Conclusão e Trabalhos Futuros | 88 |
| 7.1 | Revisão dos Objetivos e Resultados da Tese | 88 |
| 7.2 | Trabalhos Futuros | 90 |
| | Referências | 93 |
| | APÊNDICE A - Características das biosseqüências | 99 |
| | Propriedades de uma biosseqüência | 102 |
| | APÊNDICE B – Operações sobre Biosseqüências | 104 |
| | Comparação de seqüências | 104 |
| | Geração de Anotações | 106 |
| | Montagem de fragmentos | 107 |
| | APÊNDICE C – Operações sobre Biosseqüências | 110 |
| | Como as biosseqüências são armazenadas | 110 |
| | Bancos de biosseqüências | 111 |
| | Persistência em gerenciadores de bancos de dados comerciais e específicos | 121 |
| | Outras propostas de persistência para biosseqüências | 124 |
| | APÊNDICE D – Algoritmo BWT | 129 |
| | Exemplo | 130 |
| | APÊNDICE E – Implementação da solução | 131 |

Lista de Tabelas

| | |
|--|-----|
| Tabela 1 – Tipos de Programas BLAST | 28 |
| Tabela 2 – Passos do Algoritmo BLAST | 29 |
| Tabela 3 - Configuração da execução do BlastP para estudo do código. | 30 |
| Tabela 4 - Dados de entrada da execução do BLAST. | 39 |
| Tabela 5 - Dados de entrada da execução do BLAST | 40 |
| Tabela 6 - Resultados da execução do BLAST com a base nr captados pelo PFMon. | 42 |
| Tabela 7 - Resultados da execução do BLAST com a base env_nr captados pelo AQTime. | 43 |
| Tabela 8 - Comparação de programas de compressão de dados | 58 |
| Tabela 9 - Configuração de Hardware e Software Básico | 72 |
| Tabela 10 - Cenários de execução | 74 |
| Tabela 11 - Planejamento experimental | 76 |
| Tabela 12 - Configuração do testes T1 a T5 | 77 |
| Tabela 13 - Resumo dos resultados para o primeiro cenário | 79 |
| Tabela 14 - Resultados do teste T1 | 80 |
| Tabela 15 - Resultados do teste T2 | 80 |
| Tabela 16 - Resultados do teste T3 | 81 |
| Tabela 17 - Resultados do teste T4 | 81 |
| Tabela 18 - Resumo dos teste T5 | 82 |
| Tabela 19 - Configuração dos testes T6 à T10 | 82 |
| Tabela 20 - Resultados testes T6 à T10. | 84 |
| Tabela 21- Codificação para ambigüidades na leitura de nucleotídeos | 99 |
| Tabela 22 - Lista de aminoácidos | 100 |
| Tabela 23 - Matriz ordenada resultante da BWT | 130 |

Lista de Figuras

| | |
|---|-----|
| Figura 1 - Alinhamento entre seqüências | 17 |
| Figura 2 - Módulos Principais do NCBI-BLAST | 31 |
| Figura 3 - Árvore de chamada das funções do BLAST. | 33 |
| Figura 4 - Execução do PFMon para analisar page faults do BLAST | 35 |
| Figura 5 - Exemplo da informação captada com o utilitário FileMon. | 36 |
| Figura 6 - Interface do aplicativo AQTime | 37 |
| Figura 7 - Exemplo de funcionamento da primeira e segunda fase do BLAST. | 38 |
| Figura 8 - Padrão de acesso ao arquivo de seqüências em uma execução do BLAST. | 40 |
| Figura 9 - Padrão de acesso ao arquivo de seqüências em uma execução do BLAST | 41 |
| Figura 10 - Acesso aleatório ao arquivo de seqüências em uma execução do BLAST. | 42 |
| Figura 11 - Linha de código onde são acessados os dados de uma Biosseqüências | 43 |
| Figura 12 - Fórmula para Cálculo de Taxa de Compressão | 48 |
| Figura 13 - Principais Componentes da Solução Proposta | 62 |
| Figura 14 - Módulos Globais | 63 |
| Figura 15 - Compactação dos Dados de Biosseqüências | 65 |
| Figura 16 - Módulo Compactador | 66 |
| Figura 17 - Seqüência das Chamadas das Funções de Compactador | 66 |
| Figura 18 - Gerência de Memória | 67 |
| Figura 19 - Interação entre BLAST e Gerente de Memória | 69 |
| Figura 20 - Componentes do Cenário de Execução | 73 |
| Figura 21 - Número de operações de E/S para as execuções do cenário | 78 |
| Figura 22 - Tempo de execução total para as execuções do cenário 1 | 79 |
| Figura 23 - Número de operações de E/S para as execuções do cenário 2. | 83 |
| Figura 24 - Tempo total de execução para as execuções do cenário 2. | 84 |
| Figura 25 - Resultados da execução do teste T6 | 85 |
| Figura 26 - Resultados da execução do teste T7 | 85 |
| Figura 27 - Resultados da execução do teste T8 | 85 |
| Figura 28 - Resultados da execução do teste T9 | 86 |
| Figura 29 - Resultados da execução do teste T10 | 86 |
| Figura 30 - Alinhamento entre seqüências | 103 |

| | |
|---|-----|
| Figura 31 - Exemplo de intercalação na montagem de fragmentos | 108 |
| Figura 32 - Exemplo de uma seqüência guardada no DDBJ | 113 |
| Figura 33 - Arquivo ASN.1 representando uma seqüência registrada no GenBank | 115 |
| Figura 34 - Esquema de uma saída em XML obtida da execução do BLAST | 117 |
| Figura 35 - Uma seqüência em um arquivo FASTA gerado pelo NCBI | 119 |
| Figura 36 - Exemplo de um arquivo de entrada para o sistema ACeDB | 123 |
| Figura 37 - Árvore de sufixo. | 124 |
| Figura 38 - Passos do BWT | 129 |