1 Introdução

1.1 Bioinformática e Banco de Dados

Assim como o Projeto do Genoma Humano (PGH), cujas principais contribuições são a descoberta da seqüência completa de nucleotídeos do genoma humano[PS91], existem vários projetos de seqüenciamento acontecendo nos laboratórios [NP04]. Os dados gerados passam por algumas fases desde o experimento biológico em si, seqüenciamento, montagem até chegar a uma seqüência de DNA de um organismo.

Ao isolar novas seqüências em laboratório, os pesquisadores querem saber o máximo possível sobre essas seqüências. Porém, dado o tamanho das seqüências e a complexidade em sua análise, esse processo tornou-se inviável de ser realizado manualmente. Assim, vários métodos computacionais foram desenvolvidos dentro da Biologia Computacional (ou Bioinformática) para analisar seqüências genômicas e identificação dos genes.

Se a Bioinformática é uma área da ciência com muitos desafios atualmente, isto se deve, entre outros, ao volume de dados biológicos existentes e seu constante crescimento graças às técnicas de seqüenciamento avançadas, aumento do poder computacional e disponibilidade de ferramentas próprias da área.

Este grande volume de dados de biosseqüências tem forçado a busca de novas técnicas para lidar com o armazenamento e acesso eficiente aos dados. Desta maneira, é possível encontrar propostas desde otimizações das ferramentas mais utilizadas (e.g. [DBL+00]) até a extensão de gerenciadores de banco de dados relacionais [SCD+05, HPY05], além de sistemas específicos [MDS99], para resolver as principais operações da biologia computacional. É possível mencionar também outras técnicas oriundas da área de banco de dados como compactação de dados biológicos [GT93] e indexação das bases de dados da biologia [HAI01, THP04].

É importante ressaltar que usar a tecnologia de banco de dados aqui não se traduz simplesmente em usar um sistema gerenciador de banco de dados. Caso fosse adotada esta estratégia seria necessária a recodificação das aplicações que lidam com biosseqüências para acessarem um SGBD, uma vez que grande parte das aplicações que lidam com biosseqüências usam como fonte de dados arquivos texto ou arquivos semi-estrturados como o formato XML.

A idéia geral deste trabalho de pesquisa em biologia molecular é estudar o comportamento de aplicações da área em relação ao acesso aos dados de biosseqüências em memória secundária, visando a elaboração de uma proposta de persistência adequada às propriedades deste tipo de dado e à sua utilização.

1.2 Objetivos da Dissertação

Os objetivos desta dissertação são os seguintes:

1. Estudar as características das biosseqüências e os problemas associados com sua persistência em memória secundária.

2. Pesquisar sobre as técnicas de compactação de dados no domínio de sistema gerenciadores de banco de dados, avaliando as vantagens e desvantagens de sua aplicação.

3. Propor e implementar uma solução adequada para fazer uso das técnicas de compactação de dados na persistência e acesso aos dados de biosseqüências.

4. Analisar os resultados da solução implementada através da execução de diversos cenários de aplicações conhecidas da Bioinformática que façam uso intensivo dos dados de biosseqüências.

5. Análise detalhada do funcionamento do programa NCBI-BLAST com o objetivo de estudar seu comportamento durante o acesso aos dados de

biosseqüências. Esta aplicação foi escolhida por ser altamente utilizada pelos pesquisadores da área.

1.3 Estrutura da Dissertação

Após esta introdução, o texto está dividido da seguinte forma.

Preliminares: O Capítulo 2 apresenta uma discussão preliminar do contexto biológico necessário para o entendimento e motivação deste trabalho. São apresentadas as principais características dos dados de biosseqüências e a utilidade das mesmas dentro da pesquisa da biologia. Em seguida, detalha-se as formas como as biosseqüências são normalmente persistidas em memória secundária, finalizando com uma discussão selecionadas da literatura.

Análise do programa BLASTP: No Capítulo 3 é apresentada uma descrição do funcionamento do programa BLAST já que será usado como plataforma para testar algumas das idéias aqui propostas. Além disso, é realizada uma análise detalhada da versão NCBI-BLAST visando descrever seu comportamento durante o acesso aos dados de biosseqüências, identificando oportunidades de melhorar problemas relacionados com entrada/saída (E/S).

Técnicas para Compactação de Dados: O Capítulo 4 introduz os conceitos de compactação de dados, pois esta é a técnica escolhida aqui para abordar problemas levantados no capítulo 3, relacionando-os com a aplicação desta técnica dentro do contexto de banco de dados. Em seguida, faz uma revisão da literatura desta área para apresentar os trabalhos relacionados com a aplicação de técnicas de compactação sobre dados de biosseqüências.

Proposta de Solução: A proposta de solução para os problemas destacados anteriormente é apresentada no Capítulo 5. No inicio do capítulo é realizada uma breve discussão sobre a solução adotada e seus principais componentes. Finalmente, é detalhada a proposta de solução apresentando a arquitetura do software que foi desenvolvida em conjunto com as justificativas para as decisões tomadas durante seu projeto.

Análise dos Resultados: O Capítulo 6 descreve inicialmente a metodologia de testes e análise utilizada. Em seguida apresenta os resultados

obtidos discutindo cada um dos cenários que foi utilizado para a execução dos testes.

Conclusão: No Capítulo 7, têm-se os comentários finais incluindo um breve resumo da pesquisa realizada e uma enumeração das contribuições desta dissertação. Ao final, são sugeridos alguns trabalhos que poderão dar seqüência à pesquisa aqui realizada.