7 Conclusão e Trabalhos Futuros

Neste capítulo é fornecido um sumário da pesquisa realizada. Os objetivos desta dissertação, os quais foram descritos no "Capítulo 1 – Introdução", servirão como guias para avaliar os resultados obtidos. Na primeira seção são revisados os objetivos da dissertação mostrando como esses objetivos foram alcançados, e caso não tenham sido totalmente alcançados, uma justificativa é apresentada.

Ao final deste capítulo são apresentados os potenciais trabalhos futuros. Neste sentido, os trabalhos futuros são divididos em dois grupos, aqueles que se referem a presente pesquisa e aqueles que se referem à extensão desta pesquisa.

7.1 Revisão dos Objetivos e Resultados da Tese

Os cincos objetivos desta dissertação foram declarados no Capítulo 1 e são usados para estruturar a avaliação dos resultados obtidos nesta dissertação e como isto de fato foi alcançado.

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

No Capítulo 2 foi discutida a importância do uso das informações sobre biosseqüências para o avanço da pesquisa em bioinformática. Igualmente foram apresentados as principais características, os mecanismos utilizados para persistência em memória secundária e as operações mais freqüentes sobre dados de biosseqüências. Complementando esta discussão, foram apresentados os problemas encontrados com a persistência e acesso a este tipo de informação. Grande parte da informação apresentada neste capítulo pode ser complementada pelos textos presentes nos apêndices em anexo a este documento.

Ao final do Capítulo 2 foram apresentados os trabalhos relacionados com a melhoria de persistência e acesso às biosseqüências, onde foi percebida a

necessidade de adoção de uma solução que pudesse lidar com o problema da persistência, acesso e gerência de memória de forma única.

"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".

Este objetivo foi tratado no Capítulo 4, especialmente dedicado a descrever as técnicas de compactação de dados. Inicialmente foram introduzidos os conceitos básicos do termo compactação de dados. Em seguida, foram apresentados os trabalhos relacionados com as técnicas de compactação em banco de dados em geral. Finalmente, foram apresentados os trabalhos que tratavam de compactação para dados de biosseqüências. Desta forma, foi concluído que os dados de biosseqüências são candidatos ao uso das técnicas de compactação, em especial devido as suas características de alfabeto pequeno e alto grau de redundância.

"Propor e implementar uma solução adequada para fazer uso das técnicas de compactação de dados na persistência e acesso aos dados de biosseqüências. Dentro deste objetivo está incluída a análise das técnicas de gerência de memória sobre dados compactados".

No Capítulo 5 foi propostas uma solução para o problema de persistência e acesso aos dados de biosseqüências. Esta proposta foi elaborada integrando técnicas para persistência, método de acesso e gerência de memória. Desta maneira supomos que uma estratégia integrada pudesse trazer melhores resultados na solução do problema. Ainda neste capítulo foi apresentada a arquitetura de implementação da solução detalhando cada módulo e seu funcionamento.

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

O Capítulo 3 apresentou uma análise do programa NCBI-BLAST visto que ele foi utilizado como plataforma de testes para o ambiente escolhido. Para isso, foi necessário identificar os pontos críticos deste programa em termos de

acesso aos dados de biosseqüências e analisar o custo das operações de entrada e saída. Este estudo foi de grande valor para poder acoplar a solução adotada, visto que a solução apresentava um módulo para realizar a gestão de memória, a qual precisou ser integrada ao NCBI-BLAST.

"Analisar os resultados da solução implementada através da execução de diversos cenários dentro do contexto biológico. O objetivo é usar aplicações conhecidas dentro do contexto biológico, as quais façam uso intensivo dos dados de biosseqüências".

No penúltimo capítulo, Capítulo 6, foram realizados diversos testes utilizando a solução implementada. No início deste capítulo foi descrita a metodologia utilizada e apresentados os cenários interessantes para serem testados. Foram definidos dois cenários com variações no tamanho da seqüência de entrada, no tamanho do banco de dados de biosseqüências e no tamanho do bloco de dados compactado. Além disso, foram usadas três variações do NCBI-BLAST: original, com gerência de memória sem uso de compactação e com gerencia de memória usando compactação. Os resultados obtidos mostraram uma grande redução na quantidade de operações de entrada e saída nos casos de uso de compactação. Porém não aumento no tempo de execução do algoritmo NCI-BLAST, visto que o custo de processamento do algoritmo de descompactação foi muito alto. Desta maneira, foi concluído a necessidade de extensão da pesquisa no sentido de melhorar esses algoritmos para melhorar sua performance quando os dados estão em memória.

7.2 Trabalhos Futuros

Dois tipos de trabalho futuro devem ser mencionados nesta seção. O primeiro cobre as áreas nas quais devem levar a uma melhoria nos objetivos na seção anterior. O segundo trata os campos de pesquisa que poderiam estender a abrangência da pesquisa executada, e assim poderiam enriquecer a funcionalidade da arquitetura descrita acima.

Melhorias da Pesquisa Realizada

- Realizar um número maior de testes no sentido de analisar o impacto do tamanho dos blocos compactados no desempenho de execução do NCBI-BLAST.

- Especificar o modelo de custos da descompactação com objetivo de dar suporte à definição de mecanismos de otimização.

- Modificar a implementação do gerente para que ele funcione com uma *thread* e possa ser executado de forma assíncrona junto ao BLAST.

- Implementar novas estratégias de leitura e descompactação dos blocos compactados de biosseqüências. Desta forma, a estratégia pode ser escolhida de acordo com o tipo de operação que está sendo realizada.

Melhorias na Pesquisa Presente

Existem duas áreas nas quais esta pesquisa pode ser melhorada. Essas áreas estão relacionadas com a modificação dos algoritmos de compactação que trabalham em memória principal e execução de consultas compactadas.

É sabido que o uso dos algoritmos de compactação em memória causa um desempenho baixo, o que os impede de serem usados em aplicações que exijam um bom desempenho na descompactação. Desta forma, este problema deve ser investigado com intuito de tentar adaptar esses algoritmos para o contexto de biosseqüências onde o alfabeto é pequeno e existe muita redundância.

A estratégia adotada foi de descompactar em memória, outra estratégia interessante seria compactar a seqüência de entrada e realizar a comparação com os dados compactados. Neste sentido, será necessário estudar como pesquisas por similaridade podem ser realizadas sobre dados compactados.

Apesar da possibilidade de testar outros casos de da solução proposta neste trabalho, como mencionado no capítulo 6, faz-se necessário a advertência da existência de um limite de melhoras no tempo de execução que se poderá obter com o programa mesmo reduzindo o número de operações de E/S dada a arquitetura criada. Desta maneira, é sugerido também investigar o uso de propostas como em [ZHN+06], na qual é apresentada uma estrutura para utilização de compactação com banco de dados que explora a arquitetura de memória principal de uma máquina, trabalhando com dados compactados em

memória principal e fazendo a descompactação no *cache* sob demanda. Agregar esta estratégia em conjunto à arquitetura proposta poderá diminuir o tempo de execução total do programa.