UNIVERSIDADE FEDERAL DE PERNAMBUCO
GRADUAÇÃO EM ENGENHARIA DA COMPUTAÇÃO
CENTRO DE INFORMÁTICA
2016.1

TITULO: Implementação do alinhamento de proteínas em GPU utilizando OpenCL

PROPOSTA DE TRABALHO DE GRADUAÇÃO

Aluno: Jefferson Ramos Lucas dos Anjos (jrla@cin.ufpe.br)
Orientador: Edna Natividade da Silva Barros (ensb@cin.ufpe.br)

Recife, Abril de 2016.

Sumário

1. Resumo

Este trabalho apresenta uma implementação para alinhamento ótimo global de proteínas utilizando o algoritmo Needleman-Wunsch em OpenCL, bem como a sua avaliação de desempenho. A implementação proposta visa maximizar o desempenho do algoritmo fazendo uso dos recursos internos da GPU e do framework OpenCL. Como estudo de caso será utilizado o banco de dados de proteínas Swiss-Prot. Como comparativo serão utilizados softwares de terceiros, largamente utilizados na literatura, como o FASTA e o SWIPE.

2. Introdução

Nas ultimas décadas houve um aumento no descobrimento de novos genes, o que acarretou em novas pesquisas genéticas referentes à evolução molecular e detecção de doenças. Depois do descobrimento de um novo gene, os biólogos normalmente não tem ideia de sua função, sendo necessário inferir a função de um novo gene descoberto pela similaridade com um gene de função já conhecida (Durbin, et al., 1998).

Um bom exemplo são as proteínas, macromoléculas biológicas formadas por moléculas mais simples denominadas aminoácidos. Na natureza existe cerca de 300 aminoácidos, porem nas proteínas podemos encontrar 20 aminoácidos principais. As proteínas evoluem juntamente com o organismo, descobrir suas estruturas e funções em organismos vivos é importante para a compreensão dos processos celulares e a criação de novos medicamentos.

A comparação de sequências, também chamada alinhamento de cadeias, é um problema clássico em biologia molecular e resulta em uma medida de similaridade (ou distância) entre duas sequências de comprimentos arbitrários, a *query*, sequencia alvo, e o *subject*, sequencia(s) de referências. O alinhamento de sequências é dividido em duas categorias: alinhamento global, que consiste em fazer a comparação de toda a extensão das sequências e alinhamento local, que faz a busca por pequenas similaridades dentro das sequências. Alinhamento de sequências é uma abordagem comum em bioinformática para comparar genes de DNA, estudar a evolução molecular, detectar doenças, vírus e estudar a estrutura das proteínas observando padrões de conservação e variabilidade para predições estruturais e funcionais.

O algoritmo de Needleman-Wunsch (NW) é um algoritmo de alinhamento global de sequências, encontrando semelhanças por toda a extensão das sequências. O algoritmo de NW é baseado em programação dinâmica (não heurístico), esse paradigma quebra o problema original em subproblemas menores a fim de resolver esses subproblemas da melhor forma construindo uma solução ideal para o problema original (Needleman, et al., 1970).

A solução proposta neste trabalho de graduação pretende utilizar o algoritmo de NW para alinhar cadeias de proteínas em OpenCL. Muitos trabalhos também propuseram soluções em GPUs (Savran, et al., 2014), (Liu, et al., 2013) e serão utilizados como comparativos de desempenho.

3. Objetivo

O objetivo principal deste trabalho de graduação é implementar e avaliar o desempenho do alinhamento de proteínas em OpenCL utilizando o algoritmo de Needleman-Wunsch para as arquiteturas de CPU e GPU. A primeira versão será a versão canônica do algoritmo que será implementada para uso comparativo das demais. Será necessário definir uma arquitetura de software otimizada para aumentar o desempenho do algoritmo de alinhamento de sequências NW, gerar e comparar resultados com softwares da literatura, dentre eles estão o FASTA, SWIPE e a biblioteca BIOJAVA.

4. Metodologia

Este trabalho será dividido em 3 etapas principais: pesquisa, desenvolvimento e validação.

Na fase de pesquisa, serão feitos estudos aprofundados sobre alinhamento de sequências de proteínas, sua importância, o algoritmo utilizado e seus meios de otimizações. Além disso, o estudo de softwares já existentes a fim de serem usados como comparativo na fase de validação.

A fase de desenvolvimento é dividida em 2 etapas:

1. A implementação de uma versão canônica do algoritmo, sem nenhum tipo de otimização para validar a funcionalidade com os softwares de referência. Essa versão serve como base de comparação para todas as outras versões implementadas.
2. A implementação da versão em OpenCL otimizada para obter o melhor desempenho conseguido, otimizando os recursos internos das threads em OpenCL e minimizando as transferências com o Host.

A fase de validação ocorrerá mediante a obtenção dos resultados satisfatórios. Estudos comparativos serão realizados conforme a seguir:

1. Desempenho dos softwares de referência
2. Desempenho da versão canônica em CPU
3. Desempenho da versão canônica em OpenCL - CPU e GPU;
4. Desempenho da versão otimizada em OpenCL - CPU e GPU;
5. Desempenho das versões implementadas com os trabalhos relacionados.

Ao final do trabalho espera-se que haverá evidências suficientes para comprovar o desempenho da implementação em OpenCL para alinhamentos de proteínas utilizando a arquitetura de CPU e GPU.

5. Cronograma

Atividade	Março				Abril				Maio				Junho				Julho			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
Pesquisa Bibliográfica																				
Implementação do algoritmo Canônico																				
Implementação do algoritmo em OpenCL Canônico																				
Implementação otimizada em OpenCL																				
Rodar Benchmarks																				
Rodar testes e coletar resultados																				
Escrita da Monografia																				
Preparação apresentação																				

6. Referências

Durbin R. [et al.] "Biological Sequence Analysis" [Livro]. - Cambridge : Cambridge University Press, 1998.

Liu Y., Wirawan A. e Schmidt B. CUDASW++ 3.0: accelerating Smith-Waterman protein database search by coupling CPU and GPU SIMD instructions [Periódico] // BMC Bioinformatics . - 2013.

Needleman S.B. e Wunsch C.D "A general method applicable to the search for similarities in the amino acid sequence of two proteins" [Artigo] // Journal of Molecular Biology. - 1970. - Vols. vol. 48, pp. 443-453.

NIH National Human Genome Reserach Intitute - The 10-year anniversary of the Human Genome Project: commemorating and reflecting [Online] // National Human Genome Reserach Intitute. - 30 de Abril de 2013. - https://www.genome.gov/.

Savran I., Gao Yang e Bakos J.D. Large-Scale Pairwise Sequence Alignments on a Large-Scale GPU Cluster [Periódico] // Design & Test, IEEE . - 2014. - 1 : Vol. 31. - pp. 51-61 .

Setubal J. e Meidanis J. "Introduction to Computational Mocelular Biology" [Livro]. - Boston : PWS Publishing Company, 1997.

Banca:

Manoel Eusebio de Lima

Adriano Sarmento

7. Assinaturas

Aluno de Graduação: Jefferson Ramos Lucas dos Anjos

Orientadora: Edna Natividade da Silva Barros

_____

Jefferson Ramos Lucas dos Anjos
aluno

_____

Edna Natividade da Silva Barros
orientadora