Universidade Federal de Pernambuco

Centro de Informática

Graduação em Ciência da Computação

2016.1

Título: Otimização de um coprocessador paralelo de alinhamentos de DNA em FPGA

Proposta de Trabalho de Graduação

Aluno: João Gabriel Machado da Silva (jgms@cin.ufpe.br)

Orientadora: Edna Natividade da Silva Barros (ensb@cin.ufpe.br)

Recife, abril de 2016

RESUMO

A necessidade de processar as informações providas dos sequenciamentos de DNA tem crescido devido à diminuição do custo do sequenciamento. Neste trabalho é otimizado um coprocessador de alinhamento de DNA através da generalização e flexibilidade dos parâmetros da arquitetura para o alinhamento global de sequências de DNA, possibilitando assim a utilização da arquitetura em outras possíveis aplicações na área de bioinformática.

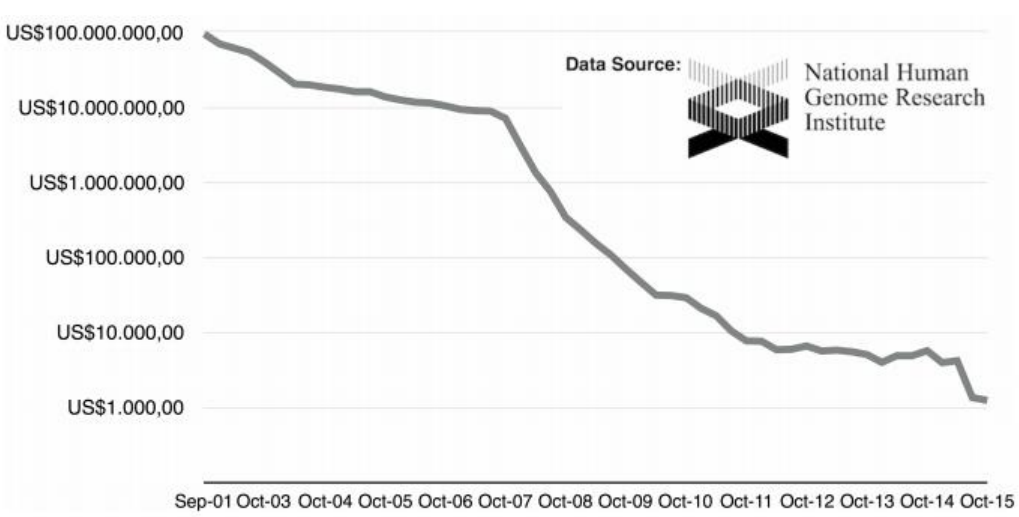
SUMÁRIO

CONTEXTO

O sequenciamento de DNA é o processo de extração da informação biológica do DNA, traduzindo-a em uma cadeia linear de símbolos. O alfabeto que compõe o DNA contem quatro símbolos, denominados nucleotídeos, são eles: a = adenina, c = citosina, t = timina, g = guanina.

Desde o primeiro sequenciamento de DNA em 1970, o custo de sequenciamento por individuo está se tornando mais barato. Por conseguinte, a quantidade de espécies completamente sequenciadas está crescendo [1]. A Figura 1 é um gráfico logaritmo que mostra a evolução do custo do sequenciamento desde 2002, quando o custo foi de US$ 100 milhões, até chegar ao ano de 2015 quando o custo aproximou-se de US$ 1 mil. Até 2007, a curva evolutiva vinha com um decaimento lento, mas com a difusão da NGS (Next Generation Sequencer) [2], a curva começou a ter um decaimento mais intenso.

Figura 1 Evolução do custo de sequenciamento de DNA



A necessidade de processar as informações providas dos sequenciamentos de DNA de modo que possa ser útil para promover avanços científicos, tem criado problemas inteiramente novos. O exemplo clássico que gerou a arquitetura atual em que esse trabalho se baseia vem de um problema da biologia molecular resolvido por um algoritmo de comparação de sequências. Em que dadas duas sequências,

se quer saber o quanto elas são semelhantes. A comparação das sequências é a operação mais primitiva usada na biologia computacional, que serve como base para muitas outras e mais complexas manipulações [1]. O processo de comparação, também chamado de alinhamento de sequências, tem como resultante uma medida de similaridade (ou distância) entre as duas sequências de comprimentos arbitrários.

A arquitetura atual implementa o algoritmo de Needleman-Wunsch que resolve o problema do alinhamento de DNA de maneira ótima (Não heurístico) e dá a medida de similaridade entre as duas sequencias alinhadas em tempo quadrático e complexidade espacial [3].

A arquitetura atual que serve como base para esse trabalho, tem como proposito fazer o alinhamento de DNA de indivíduos humanos, tendo como motivação a busca em grandes bancos de sequências forense, auxiliando as autoridades na resolução de crimes de investigação. Nos EUA o sistema chamado CODIS (Combined DNA Index System) foi criado em 2007. Este sistema é um programa do FBI para dá suporte a justiça criminal na base de dados de DNA. Atualmente, o banco de dados da CODIS conta com quase 14,5 milhões de perfis de DNA, incluindo criminosos detidos e perfis forenses. Até junho de 2015, o sistema produziu mais de 288.298 visitas de assistência em mais de 274.648 investigações [4].

Neste contexto, o problema de verificação da identidade do criminoso consiste em comparar o material recolhido na cena do crime com todos os perfis existentes armazenados no banco de dados, em um estilo de consulta um pra n. Além disso, cada indivíduo humano é identificado por *p* sequencias. Na base de dados do Reino Unido um individuo humano é composto por p = 15 sequências, no entanto na CODIS é composto por p = 13.

A arquitetura atual foi elaborada fixando o comprimento máximo de cada sequência do individuo, e a quantidade de PE (Processor Elements) utilizado para computar uma sequência, como também foi fixada a quantidade p = 15 de sequências por individuo humano e a quantidade de indivíduos completos que a arquitetura computa em paralelo, a parametrização da arquitetura contem restrições

que impendem a generalização e flexibilidade da arquitetura para outras aplicações, restringindo a arquitetura apenas para a resolução dessa única aplicação forense.

A outras aplicações onde é possível utilizar a arquitetura atual, tais como taxonomia, previsão da estrutura das proteínas, identificação pessoal, engenharia genética, e muitos outros [1].

OBJETIVO

Motivado pelas possíveis aplicações para a área de bioinformática, propõe-se aumentar a parametrização da arquitetura atual. Possíveis otimizações no uso de recursos e da frequência máxima da arquitetura atual, também pretende-se investigar.

Esse trabalho tem como objetivo principal, otimizar um coprocessador de alinhamento de DNA através da generalização e flexibilidade dos parâmetros da arquitetura para o alinhamento global de sequências de DNA, como tamanho da sequência, número de PE (Processor Elements), quantidade de sequências por Indivíduos e indivíduos em paralelo. Adicionalmente, esse trabalho visa propor melhorias na arquitetura de forma a torna-la escalável para alguns parâmetros.

METODOLOGIA

Será realizado um estudo da arquitetura atual a fim desse autor conhecer os conceitos empregados na mesma, elucidando assim toda a parametrização atual da arquitetura, buscando pontos de ajuste na parametrização.

Após esse estudo, será feito os ajustes necessários na parametrização da arquitetura, para que a mesma possa ser funcional e permita uma generalização da arquitetura para outras configurações e aplicações. Durante esse processo de ajuste será possível fazer uma analise mais profunda da arquitetura encontrando pontos de otimização, assim será proposto mudanças na arquitetura visando sua otimização tanto no uso de recursos, quanto da frequência máxima.

Seguindo o fluxo de execução desse trabalho, conseguinte à proposição de melhorias será feito a implementação das mesmas. Durante todo o processo já descrito, será feito a verificação funcional dos ajustes feitos e teste e validação dos resultados obtidos com as proposições de melhorias. Para que realização dos testes seja feita é preciso fazer ajustes na aplicação de prototipação da arquitetura.

As atividades acontecem em paralelo pela necessidade que as convergem. Para dar início a escrita da monografia é preciso fazer a analise bibliográfica desse trabalho para conhecer melhor o estado da arte da área de aplicação da arquitetura.

Por fim, ao termino da escrita da monográfica será feito a apresentação desse trabalho de graduação para a defesa desse trabalho.

CRONOGRAMA

O cronograma previsto para a execução desse trabalho de graduação pode ser visualizado na Tabela 1.

Tabela 1 Cronograma

Atividade	Março				Abril				Maio				Junho				Julho			
Estudo da Arquitetura																				
Ajuste da Parametrização																				
Analise e Proposição de mudanças																				
Implementação das mudanças																				
Verificação Funcional																				
Ajustes na Aplicação																				
Teste e Validação dos resultados																				
Analise bibliográfica																				
Escrita da Monografia																				
Preparação da Apresentação																				

REFERÊNCIAS

[1] J. Setubal e J. Meidanis, Introduction to computational molecular biology, PWS Publishing, 1997.

[2] J. Xu, Next Generation Sequencing: Current Technologies and Applications, Caister Academic Press, 2014.

[3] S. Needleman e C. Wunsch, “A general method applicable to the search for similarities in the amino acid sequence of two proteins,” *Journal of Molecular Biology,* vol. 48, n. 3, pp. 443-453, Março 1970.

[4] “CODIS - NDIS Statistics,” FBI, [Online]. Available: https://www.fbi.gov/about-us/lab/biometric-analysis/codis/ndis-statistics. [Acesso em 13 de Abril de 2016].

POSSÍVEIS AVALIADORES

1º Professor Manoel Eusébio de Lima (mel@cin.ufpe.br)

2º Professor Abel Guilhermino da Silva Filho (agsf@cin.ufpe.br)

ASSINATURA

Título: Otimização de um coprocessador paralelo de alinhamentos de DNA em FPGA

Aluno: João Gabriel Machado da Silva

Orientadora: Edna Natividade da Silva Barros

Edna Natividade da Silva Barros
(Orientadora)

João Gabriel Machado da Silva
(Aluno)