KU LEUVEN

FACULTEIT INDUSTRIËLE
INGENIEURSWETENSCHAPPEN

CAMPUS BRUGGE

Genetic sequence alignment acceleration using a FPGA based platform

What methods can be used to accelerate the Smith-Waterman algorithm for genetic sequence alignment with an FPGA equipped platform?

Robin NOLLET

Supervisors: Ing. Václav Šimek
: Ing. Jonas Lannoo

Master Thesis to obtain the degree of
Master of Science in Engineering Technology:
Electronical engineering

Academic Year 2019 - 2020

Acknowledgements

Het voorwoord vul je persoonlijk in met een appreciatie of dankbetuiging aan de mensen die je hebben bijgestaan tijdens het verwezenlijken van je masterproef en je hebben gesteund tijdens je studie.

Summary

De (korte) samenvatting, toegankelijk voor een breed publiek, wordt in het Nederlands geschreven en bevat **maximum 3500 tekens**. Deze samenvatting moet ook verplicht opgeladen worden in KU Loket.

Abstract

Het extended abstract of de wetenschappelijke samenvatting wordt in het Engels geschreven en bevat **500 tot 1.500 woorden**. Dit abstract moet **niet** in KU Loket opgeladen worden (vanwege de beperkte beschikbare ruimte daar).

Keywords: Voeg een vijftal keywords in (bv: Latex-template, thesis, ...)

Contents

List of Figures

List of Tables

List of symbols

Maak een lijst van de gebruikte symbolen. Geef het symbool, naam en eenheid. Gebruik steeds SI-eenheden en gebruik de symbolen en namen zoals deze voorkomen in de hedendaagse literatuur en normen. De symbolen worden alfabetisch gerangschikt in opeenvolgende lijsten: kleine letters, hoofdletters, Griekse kleine letters, Griekse hoofdletters. Onderstaande tabel geeft het format dat kan ingevuld en uitgebreid worden. Wanneer het symbool een eerste maal in de tekst of in een formule wordt gebruikt, moet het symbool verklaard worden. Verwijder deze tekst wanneer je je thesis maakt.

b	Breedte	$[mm]$
A	Oppervlakte van de dwarsdoorsnede	$[mm^2]$
c	Lichtsnelheid	$[m/s]$

List of abbreviations

Chapter 1

Introduction

Chapter 2

Introduction to bioinformatics and embedded systems

2.1 Introduction to bioinformatics

2.1.1 Biology and DNA

DNA, or Deoxyrebonucleic Acid, is what stores the genetic information of all living organisms. It is the information that programs all of the activities in a cell.

2.1.1.1 Structure of DNA

Structurally, DNA is a polymer, which means each molecule is built up out of small repeating molecular units. In DNA, these units are called *nucleotides*.

Each nucleotide consists of 3 parts:

1. A carbon sugar molecule called *Deoxyribose*.
2. A phosphate group to connect the Deoxyribose molecules with eachother.
3. One of four possible nitrogen bases: Adenine (A), Thymine (T), Cytosine (C) or Guanine (G)

It is important to note that in most living organisms DNA does not exist as a single polymer, but rather a pair of molecules that are held thightly together. This is the famuous *double helix*.

Like in any good structure, there is a need for a main support. In DNA, the sugars and phosphates bond together to form twin backbones. These sugar-phosphate bonds run down each side of the helix, but chemically in opposite directions.

The first phosphate group, at the start of the molecule, connects to the sugar group's 5th carbon. At the end of the structure the 3rd carbon of the sugar group is unconnected. This makes a pattern typically noted as $[5' \rightarrow 3']$. Now, since the other molecule in the helix goes in the opposite direction, the pattern of the other backbone is typically noted as $[3' \rightarrow 5']$.

These two long chaines are linked together by the nitrogen bases via their relatively weak hydrogenbonds, but there can't just be any pair of nitrogen bases. Adenine can only make hydrogen bonds with Thymine, likewise Guanine can only bond with Cytosine. These bonded nitrogenbases are called *base paires*

It is the order of these bases, which is also called the *sequence*, that allows this DNA to store usefull information. In this way, e.g. $AGGTCCATG$ means something completely different as a base sequence than e.g. $TTCCAGATC$.

Since each of the bases in the sequence has only one possible counterpart, you are able to predict what its matching counterpart will be in the opposite string. For example:

If the following sequence is known

$$[5' - AGGTCCG - 3']$$

we can deduce the sequence in the other direction as

$$[3' - TCCAGGC - 5']$$

2.1.1.2 DNA in the human body

In human cells, DNA molecules can be found in the nucleus of all cells in the body. It consists of 46 very long molecules, which during cell division condense in what we call *chromosomes*. The only exception is reproductive cells, which only has 23 chromosomes. These chromosomes are packed tighly together in the nucleus of the cell. If all of these chromosomes are put together, this makes about 3 billion base pairs. These 3 billion base pairs provides the assembly instructions for pretty much everything inside the cell.

These 46 chromosomes, which make up our whole DNA, are always present in pairs in the cells. Each time, the pair consists of one chromosome from each parent.

These 23 chromosome pairs are classified in

- 2 times 22 automomal chromosomes. These are marked 1 to 22 according to the length of the sequence. The longest chromosome (chromosome number-1) is 248,956,422 bases long. The shortest (chromosome number-22) is 50,818,468 bases long.
- In each cell there is also an X chromosome plus an X or Y, dependent on the gender.

2.1.2 Sequencing and the need for Bioinformatics

2.1.3 DNA sequence aligning

2.1.4 Clinical applications

2.2 Platforms for sequence alignment algorithms

2.2.1 CPU

2.2.2 GPU

2.2.3 FPGA

2.2.4 ASIC

2.3 Problem definition

Chapter 3

Methods for genetic sequence alignment

3.1 Local VS global alignment

3.2 commonly used algorithms

3.2.1 Dynamic programming algorithms

3.2.1.1 Needleman-Wunsch

3.2.1.2 Smith-Waterman

3.2.2 Heuristic algorithms

3.2.2.1 FASTA

3.2.2.2 BLAST

3.3 algorithm selection

3.3.1 Smith Waterman

Chapter 4

Reference mapping accelerated

4.1 problems with the direct approach

4.2 acceleration techniques

Chapter 5

System implementation for reference genome mapping

...

Chapter 6

implementation results and speedup

...

Chapter 7

Conclusion and future research

...

Bibliography

Aad, G., Abajyan, T., Abbott, B., Abdallah, J., Khalek, S. A., Abdelalim, A., Abdinov, O., Aben, R., Abi, B., Abolins, M., et al. (2012). Observation of a new particle in the search for the standard model higgs boson with the atlas detector at the lhc. *Physics Letters B*, 716(1):1–29.

Cottrell, J. A., Hughes, T. J., and Bazilevs, Y. (2009). *Isogeometric analysis: toward integration of CAD and FEA*. John Wiley & Sons.

Hughes, T. J., Cottrell, J. A., and Bazilevs, Y. (2005). Isogeometric analysis: Cad, finite elements, nurbs, exact geometry and mesh refinement. *Computer methods in applied mechanics and engineering*, 194(39):4135–4195.

Appendix A

Uitleg over de appendices

Bijlagen worden bij voorkeur enkel elektronisch ter beschikking gesteld. Indien essentieel kunnen in overleg met de promotor bijlagen in de scriptie opgenomen worden of als apart boekdeel voorzien worden.

Er wordt wel steeds een lijst met vermelding van alle bijlagen opgenomen in de scriptie. Bijlagen worden genummerd het een drukletter A, B, C,...

Voorbeelden van bijlagen:

Bijlage A: Detailtekeningen van de proefopstelling

Bijlage B: Meetgegevens (op USB)

FACULTEIT INDUSTRIËLE INGENIEURSWETENSCHAPPEN
CAMPUS BRUGGE
Spoorwegstraat 12
8200 BRUGGE, België
tel. + 32 50 66 48 00
iiw.brugge@kuleuven.be
www.iiw.kuleuven.be