

GPU Accelerated Pattern Matching Algorithm for DNA Sequences to Detect Cancer using CUDA

Dissertation

Submitted in partial fulfillment of the requirement

for the degree of

Master of Technology in Computer Engineering

By

Snehal P. Adey

MIS No: 121122001

Under the guidance of

Dr. Vandana Inamdar



**Department of Computer Engineering and Information Technology
College of Engineering, Pune**

Pune - 411005

June, 2013

DEPARTMENT OF COMPUTER ENGINEERING AND
INFORMATION TECHNOLOGY,
COLLEGE OF ENGINEERING, PUNE

CERTIFICATE

This is to certify that the dissertation titled

**GPU Accelerated Pattern matching algorithm for DNA sequences to
detect cancer using CUDA**

has been successfully completed

By

Snehal P. Adey

121122001

And is approved for the partial fulfillment of the requirements for the degree of
Master of Technology, Computer Engineering

Dr. Vandana Inamdar
Project Guide,
Department of Computer Engineering
and Information Technology,
College of Engineering, Pune,
Shivaji Nagar, Pune-411005.

Dr. J. V. Aghav
Head,
Department of Computer Engineering
and Information Technology,
College of Engineering, Pune,
Shivaji Nagar, Pune-411005.

Date _____

Abstract

The objective of string matching algorithm is to locate the appearance of a specific pattern in an array of equal or larger size text. String matching algorithms has been used in many applications such as DNA analysis. This report introduces an accelerated approach of string matching algorithm to detect the occurrence of several gene patterns in the human DNA sequence and verify whether the person has chances of getting cancer or not. DNA is a large database, an efficient algorithms is required to carry out the cancer diagnosis. This can be accomplished by parallelization technique on GPU using CUDA programming model.

Acknowledgments

I express my sincere gratitude towards my guide Dr. Vandana Inamdar for her constant help, encouragement and inspiration throughout the project work also for providing me infrastructural facilities to work in. Without her invaluable guidance, this work would never have been a successful one. I am extremely thankful to our head Dr. J. V. Aghav for all necessary cooperation in the accomplishment of dissertation. Last but not least, I would like to thank my family and friends, who have been a source of encouragement and inspiration throughout the duration of the project.

Snehal Panjabrao Adey
College of Engineering Pune

List of Symbols

DNA	Deoxyribonucleic acid
A	Adenine
C	Cytosine
T	Thymine
G	Guanine
SNP	Single nucleotide polymorphism
IARC	International Agency for research on cancer
WHO	World Health Organization
RET	RET proto-oncogene rearranged during transfection
TP53	Tumor protein p53
BRCA1	Familial breast cancer gene 1
BRCA2	Familial breast cancer gene 2
APC	Adenomatous polyposis of the colon gene
GPU	Graphics processing unit
CUDA	Compute unified Device Architecture
PCI	Peripheral Component Interconnect
AGP	Accelerated Graphics Port
API	Application programming interface
MSMPMA	Multiple Skip Multiple Pattern Matching algorithms
IBKPMPM	An Index Based K Partition Multiple Pattern Matching Algorithm
EPMSPP	Exact Multiple Pattern Matching Algorithm using DNA Sequence and Pattern Pair
IAEMA	An Intelligent and Efficient Matching Algorithm to Finding a DNA Pattern
NFA	Non-Deterministic finite Automata
DFA	Deterministic finite Automata

Contents

Abstract	I
Acknowledgement	II
List of Symbols	III
List of Figures	IV
1 Introduction	1
1.1 Bioinformatics Background.....	2
1.1.1 DNA.....	2
1.1.2 DNA Sequencing.....	3
1.1.3 Genome Sequence.....	4
1.2 Cancer Statistics Worldwide.....	5
1.2.1 Cancer Survey.....	5
1.2.2 Cancer Diagnosis Techniques.....	6
1.2.3 Cancer Genes.....	8
1.3 Parallel Programming Model.....	9
1.3.1 GPU.....	9
1.3.2 CUDA.....	11
1.4 Project Objective.....	14
1.4.1 Objectives.....	14

2	Literature Survey	15
2.1	Study of string matching algorithms.....	15
2.1.1	Traditional matching Algorithms.....	15
2.1.2	Enhanced Pattern Matching techniques.....	17
2.2	Pros and cons of matching techniques.....	19
3	Design Model	20
3.1	CODE flow.....	20
4	Implementation and Results	22
4.1	System Requirement.....	23
4.2	Results.....	23
4.2.1	Time versus Number of Patterns.....	23
4.2.2	Time versus Chunk Size.....	25
5	Conclusion and Future Scope	28
5.1	Application.....	28
5.2	Conclusion.....	28
5.3	Future Scope.....	28
6	Publication	29
7	Bibliography	30

List of figures

1.1	Structure of DNA.....	3
1.2	Estimated Cancer Deaths in the US in 2013[18].....	6
1.3	Cancers with their responsible genes. (With reference to cancer_gene_census [20]).....	9
1.4	GPU Architecture.....	10
1.5	Basic CUDA Architecture.....	12
1.6	CUDA programming model.....	13
2.1	NFA for ACATG.....	16
2.2	Deterministic Finite Automata for ACATG.....	16
2.3	Algorithm with their pros and cons.....	19
3.1	Code flow of the Pattern matching.....	21
4.1	Graph showing Processing times for parallel and serial implementation with increasing patterns (with reference to Table 4.1).....	24
4.2	Time versus No of Patterns (with reference to Table 4.2).....	25
4.3	Graph showing Processing times for parallel and serial implementation with increasing chunk size(with reference to Table 4.3).....	26
4.4	Figure 4.3 Resultant output of the pattern matching algorithm for DNA sequence.....	27

Chapter 1

Introduction

DNA is the hereditary material in almost all organisms. DNA sequence refers to the combination of A, C, T, G bases which are located on DNA strand. Gene is the basic biological unit of the DNA. It's a part of DNA which signifies organism's physical characteristics, cause of getting certain disease. In humans, all cancers arise due to mutation in the gene. An altered gene is nothing but a small change in the part of DNA. Basically gene mutation categorizes into two types, germline mutation and acquired. When the gene is passed directly from the parents to child, it is called germline mutation. Acquired mutations are those which are caused due to some factors like tobacco, (UV) radiations, viruses etc. Mutation can be harmful or beneficial, or doesn't have any effect. It depends on where the gene mutation occurs. It takes very long to get cancer from a small mutation, that is why, cancer is seen often in old people where multiple mutation happens as time passes. So early stage detection of single mutation will help to prevent further chances of getting cancer by having proper treatment and avoiding the factors that can lead further changes in the gene. This research has undergone various cancer research projects, and molecular science. The study has been done on the process of cancer detection techniques and introduced an accelerated the pattern matching process that can be useful in the medical diagnosis of cancer up to some extent.

1.1 Bioinformatics Background

Bioinformatics field has taken a vast place in the world of science as it tells about life system of the organisms. Genome analysis that plays with DNA is becoming the core interest of the scientists. It is the biological information of organism which is handled by the informative techniques with the knowledge computer science. In other words, it is a field of science which applies computer science and information technology to the problems of biological science. Biological science of an organism includes its structural, functional as well as molecular specifications. Molecular specification carries the information about the smallest unit of organism. Being the smallest this unit plays the major role in overall functioning of organism. This unit is none but a DNA.

1.1.1 DNA

DNA is the hereditary material in almost all organisms. DNA resides in every cell in the body of organism. The DNA is a double helix like structure made up of two twisted strands. A single strand carries a nucleotide bases, these bases are adenine (A), guanine (G), cytosine (C), thymine (T). These chemical letters are in the form of triplets which defines a meaningful code. The triplets give rise to specific amino acid which later helps in the formation of protein. An important property of DNA is that it can replicate, or make copies of itself. Each strand of DNA in the double helix can serve as a pattern for duplicating the sequence of bases. This is critical when cells divide because each new cell needs to have an exact copy of the DNA present in the old cell.

Functionally, DNA carries the information needed to construct and operate an organism. Other sections of DNA are responsible for switching genes on and off and regulating how much of each type of protein is made.

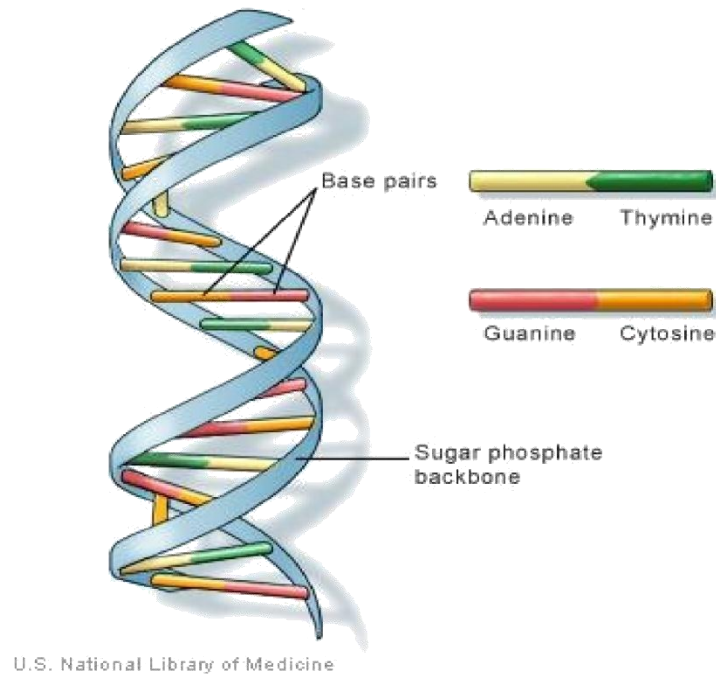


Figure 1.1 Structure of DNA [7]

1.1.2 DNA Sequencing

The term DNA sequencing refers to the combination of A, C, T, and G letters as per their arrangements in DNA. DNA sequencing is used to determine the sequence of individual genes, larger genetic regions, full chromosomes or entire genomes. To get DNA sequenced it should undergo complex laboratory computation process.

Scientists have taken a lot of efforts in the computation of DNA; it's a bit complex process needs so much accuracy. Till today the status of DNA computation is not fully worked out. DNA computations have not yet exceeded the power of modern computers.

As per the DNA computing model which was developed by a team run by Adleman in 1994, the process followed various steps. These processes are standard laboratory procedures used by biologists in performing genetic analysis [7].

After the final step of detection and reading we get sequenced DNA in a solution, then the order of sequence of base pairs that compose the strand of DNA is determined.

1.1.3 Genome Sequencing

The biological information is acquired by gathering, storing, analyzing and integrating the biological structure of organisms that brings out genetic information. It is necessary to use this genomic information in understanding human diseases and in the identification of new molecular targets for discovery. Genome sequencing is a laboratory process that determines the complete DNA sequence of an organism's genome at a single time. This entails sequencing all of an organism's chromosomal DNA as well as DNA contained in the mitochondria and, for plants, in the chloroplast. Almost any biological sample containing a full copy of the DNA, even a very small amount of DNA or ancient DNA can provide the genetic material necessary for full genome sequencing. Such samples may include saliva, epithelial cells, bone marrow, hair, seeds, plant leaves, or anything else that has DNA-containing cells. Because the sequence data that is produced can be quite large (for example, there are approximately six billion base pairs in each human diploid genome). Genomic data is stored electronically and requires a large amount of computing power and storage capacity. Full genome sequencing would have been nearly impossible before the advent of the microprocessor, computers, and the Information Age. Unlike full genome sequencing, DNA profiling only determines the likelihood that genetic material came from a particular individual or group; it does not contain additional information on genetic relationships, origin or susceptibility to specific disease. Also unlike full genome sequencing, SNP genotyping covers less than 0.1% of the genome. Almost all truly complete genomes are of microbes; the term "full genome" is thus sometimes used loosely to mean "greater than 95%". In general, knowing the complete DNA sequence of an individual's genome does not, on its own, provide useful clinical information, but this may change over time as a large number of scientific studies continue to be published detailing clear associations between specific genetic variants and diseases[5].

1.2 Cancer Statistics Worldwide

Cancer is responsible for one in eight deaths worldwide. It encompasses more than 100 distinct diseases with diverse risk factors and epidemiology which originate from most of the cell types and organs of the human body. All cancers occur due to mutation in the gene sequence.

1.2.1 Cancer Survey

The statistical data disclosed by the UK research center and International Agency for research on cancer (IARC) addresses that India has lower cancer rates than foreign countries.

- Survey in UK

UK research Centre has surveyed cancer deaths in WHO regions of the world which include Western Pacific, Europe, The Americas, South East Asia, Eastern Mediterranean and Africa and revealed that 1,57,250 died of cancer which equates almost 253 deaths for every 1,00,000 people. The study of cancer research in 2010 throughout the UK showed up 28% cancer death rate with 77,400 more deaths than coronary heart diseases. It has been detected that percentage of mortality is found to be more in males than in females. Three-quarters of deaths at age of 65 and over and half are found at the age 75 and above. The result revealed lung, bowel, breast and prostate cancer being the most common types of cancer which caused maximum no of deaths. Lung cancer is responsible for maximum number of killings attributed to smoking. Still it is estimated that 57% of the cancers depend on the genetic factors and rest depend on environmental conditions as well as changing life style.

Table 1.1 is the estimated percentage of cancer amongst male and female [18]. The data has been collected by American research Center.

Cancer	Men	Cancer	Women
Lung & Bronchus	28%	Lung & Bronchus	26%
Prostate	10%	Breast	14%
Colon & rectum	9%	Colon & rectum	9%
Pancreas	6%	pancreas	7%
Liver & intrahepatic bile duct	5%	Ovary	5%
Leukaemia	4%	Leukaemia	4%
Esophagus	4%	Non-Hodgkin lymphoma	3%
Urinary Bladder	4%	Uterine corpus	3%
Non- Hodgkin lymphoma	3%	Liver & intrahepatic bile duct	2%
Kidney & renal pelvis	3%	Brain/other nervous system	2%

Table 1.1 Estimated Cancer Deaths in the US in 2013[18]

- Survey in India

India's cancer statistic was introduced by IARC, In India death count reached to 555000 caused by cancer in 2010. Total 7137 deaths happened due to cancer in a count of 122,429 deaths, which is near about 6%. In this 71% population is recognized with age group of 30-69. The latest ever global "Novartis International AG" surveyed 1273 women in 12 countries with advanced breast cancer (Stage IV)[17].

Oral and lung cancer appear to be the mostly occurring cancers in India. Cervical, Stomach and breast accounted to be in all 41% deaths in women in urban and rural areas. Early detection and treatment of cervical, oral and breast can reduce the cancer mortality rate. Tobacco related cancers represented 42% deaths in male and 18% in females.

1.2.2 Cancer Diagnosis Techniques

There are various cancer detection techniques in existence. Lots of new methods are being developed and few more are still in process. But none of the tests will give accurate results. This is because of the technology limitation, machine error or human error. The most common cancer diagnostic methods are Biopsy, Endoscopy, Diagnostic imaging, Blood test, Pap test and genetic test.

In Biopsy, a very small portion of tissue is taken from suspected cancer cell with the help of a fine tipped needle or thicker hollow needle, or by surgical excision. This sample is then sent to laboratory for examination. Endoscopy comes with a thin flexible tube with a tiny camera on the tip of it. This tube is inserted into the body cavity. This allows doctors to view the suspicious area. This test is best suited for detecting the problems of esophagus, stomach and first part of small intestine, usually detects the Gerd, indigestion. This can be found even without endoscopy. The tube insertion may have risk of causing other damage to intestines, may need surgery in many cases. Diagnostic imaging includes X-rays, CT scan, MRI scans of various parts of the body. But this text can cause health harm due to exposure to radiation.

One more cancer detecting method which is being up for years, it is a Genetic testing. In simple words, Genetic testing is a reading a DNA code to identify abnormalities. Genome sequencing project has brought up new discovery in the DNA computation, which has made one more way to analyze disease strategy. In real it's rather a slow process of gene scanning but has tremendous impact in the molecular science of organism. Due to this it's become little bit easy to analyze, manipulate and modify gene. Such information is useful in revealing human problems. Gene testing is carried out with different techniques. One of the popular methodologies is FISH (Fluorescence in Situ Hybridization), in which specimen is scanned under fluorescent lighting. This is done in two steps. First a short sequence of single stranded DNA is prepared to match part of gene called as probe. Latter these probes get labeled with fluorescent dye. The fluorescent color will directly detect the part where actually mutation happened in the sequence.

- Genetic Testing

Gene is the basic biological unit of the DNA. It's a part of DNA which signifies organism's physical characteristics, cause of getting certain disease. In humans, all cancers arise due to mutation in the gene. An altered gene is nothing but a small change in the part of DNA. Basically gene mutation categorizes into two types, Germline and Acquired mutation. When the gene is passed directly from the parents to child, it is called Germline Mutation. Acquired mutations are those which are caused due to some factors like tobacco, (UV) radiations, viruses etc. Mutation can be harmful or beneficial, or doesn't have any effect. It depends on where the gene mutation occurs. It takes very long to get cancer from a small mutation. That is why; cancer is seen often in old people where multiple mutations happen as time passes. So early stage detection of single mutation will help to prevent further chances of getting cancer by having proper treatment and avoiding the factors that can lead further changes in the gene.

This actually undergoes two stages. In which, a blood specimen is to be tested under complex laboratory procedures so as to detect which gene is affected with a mutation and can lead to cancer. In first stage searching of mutation in gene is just like to find a single spelling mistake in the whole book of sequence. If a gene is identified to increase the risk of cancer in first stage, the affected person is suggested to have screening or predictive or predisposition testing in second stage named predictive testing. This test is kind of confirmation test which involves further testing which tells whether person is inherited mutation in cancer susceptibility gene, which increases the risk of specific cancer.

1.2.3 Cancer Gene

There are main two types of genes that play role in the cause of cancer. Those are oncogenes and tumor suppresser genes.

Oncogenes: - Most oncogenes are mutations of certain normal genes called proto-oncogenes. Proto-oncogenes controls the division of the cell so are the good ones. When a proto-oncogene mutates into an oncogene, it becomes a "bad" gene that results into activation which is not feasible. When this happens, the cell grows out of control, which can lead to cancer. E.g. RET.

Tumor Suppressor genes: - Tumor suppressor genes are normal genes that slow down cell division, repair DNA mistakes and it carries out apoptosis (programmed cell death). When tumor suppressor genes don't work properly, cells can grow out of control, which can lead to cancer. Many different tumor suppressor genes have been found, including TP53 (p53), BRCA1, BRCA2, APC, and RB1.

DNA mismatch repair gene: - Another type of genes maintains the integrity of genes and provide the accuracy in the information transfer from one gene to another. When such kind of gene gets mutated due to some abnormality, it leads to infeasibility in the cell functioning. XRCC3 gene being a DNA mismatch repair gene responsible for skin cancer.

The genes responsible for specific type of cancer are listed in the Table 1.2.

Cancer Types	Responsible Gene	Gene Location	Chromosome sequence
Lung	T53	17p13.1	Chromosome: 17;
Breast	HER2,BRCA1/2,AKT 1	17q12,17q21, 14q32	Chromosome: 17, Chromosome: 17, Chromosome: 14
Prostate	HPC 1, ACSL3, BRCA1, BRCA2, CANT1	C15orf21, 2q36 17q25	Chromosome 15, Chromosome 2, Chromosome 17
Pancreas	AKT 2, APC	19q13, 5q21	Chromosome 19, Chromosome 5
Ovary	AKT 1, AKT 2, BRCA 1, BRCA 2	14q32, 19q13	Chromosome 14, Chromosome 19
Colon & rectum	APC, ARID1A , FAM123B/WTX, TP53, SMAD4, PIK3CA and KRAS	5q21, 1p35.3, Xq11.1, 17p13.1, 3q26.3, 12p12	Chromosome 5, Chromosome 1, Chromosome x, Chromosome 17, Chromosome 3, Chromosome 12
Acute Lymphoblastic Leukaemia	MLLT2, MYC, ZNFN1A1,LAF4	4q21, 8q24.12-24.13, 7p13-p11.1	Chromosome 4, Chromosome 8, Chromosome 7,

Table 1.2 Cancers with their responsible genes. (with reference to cancer_gene_census [20])

1.3 Parallel Programming Model

1.3.1 GPU

- History of GPU

1st GPU came in 1999-2000 by NVIDIA. Many more scientific applications have been accelerated by GPU. Almost all computational vendors started adopting this new technology as it is being providing high end services and improved industry standards.

Graphics chips designed by NVIDIA constrained for specific set of functionality, it serves highly parallel programming and computation environment. This way in 1999-2000 timeframe, domain scientists and computer experts have been started using GPU's for accelerating the scientific applications and achieved the efficient performance gain. This was the advent of the movement called GPGPU, or General-Purpose computation on GPU. GPU computing is the use of a GPU together with a CPU to accelerate general-purpose scientific and engineering applications.

Pioneered five years ago by NVIDIA, GPU computing has quickly become an industry standard, enjoyed by millions of users worldwide and adopted by virtually all computing vendors [7]. From a user's perspective, applications simply run significantly faster.

CPU + GPU is a powerful combination because CPUs consist of a few cores optimized for serial processing, while GPUs consist of thousands of smaller, more efficient cores designed for parallel performance. Serial portions of the code run on the CPU while parallel portions run on the GPU.

- GPU Architecture

GPU architecture is built with a specialized circuit which could accelerate the output image in a frame buffer intended for output to display. GPU's are very efficient at manipulating computer graphics and are generally more effective than general purpose CPU's for algorithms where processing of large blocks of data is done in parallel.

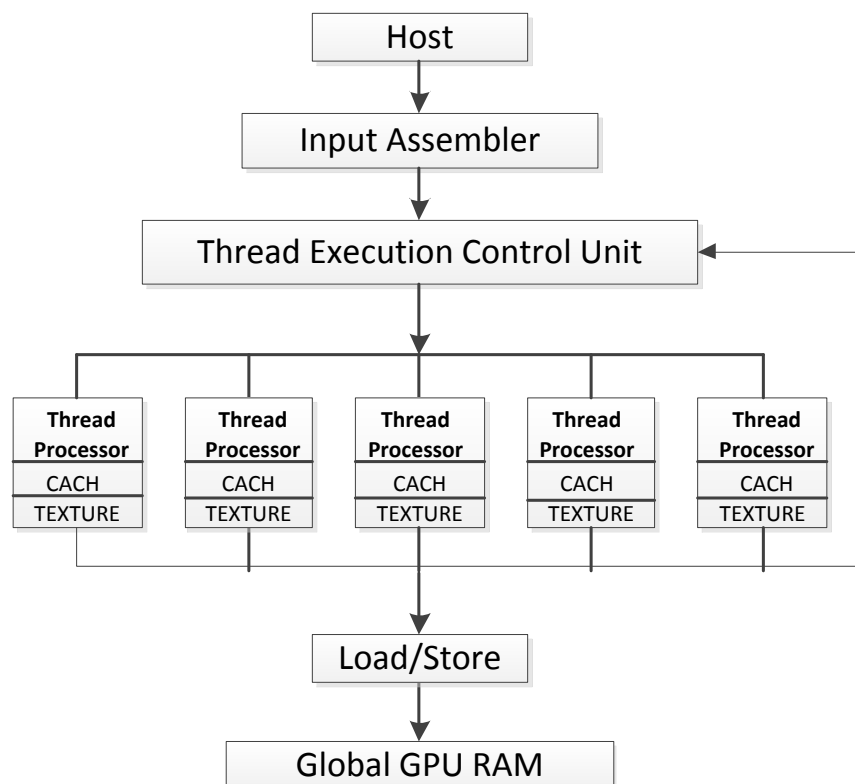


Figure 1.2 Basic GPU Architecture

A GPU is tailored for highly parallel operation while a CPU executes programs serially. For this reason GPU has many parallel execution units and higher transistor counts as compared to CPU which few execution counts against the higher clock speeds. GPU's have much deeper pipelines. This pipeline proceeds with host interface and ends at memory interface. Host interface is followed by vertex processing later sets up the triangle finally pixel processing is done before memory interface.

In a modern architectural systems CPU is connected with GPU via PCI Express or AGP slots being a graphics connector on the motherboard in order to communicate with GPU resources. Graphics connector transfers all commands, texture, and vertex data from CPU to GPU. The graphics connector is being improved and introduced as bus technology. Previously AGP slots being 32 bit wide would run at 66 MHz bandwidth and process the data at 264 Mb/sec. Later AGP came up with new releases followed 2x, 4x, 8x which ultimately doubling the bandwidth. This continued till PCI expressed standard has introduced in 2004 processing with approximately 4 Gb/sec of bandwidth to and from the GPU [8].

- GPU Computing

At the start of multicore CPUs and GPUs the processor chips have become parallel systems. But speed of the program will be increased if software exploits parallelism provided by the underlying multiprocessor architecture [1]. Hence there is a big need to design and develop the software so that it uses multithreading, each thread running concurrently on a processor, potentially increasing the speed of the program dramatically. To develop such a scalable parallel applications, a parallel programming model is required that supports parallel multicore programming environment.

NVIDIA's graphics processing units (GPUs) are very powerful and highly parallel. GPUs have hundreds of processor cores and thousands of threads running concurrently on these cores, thus because of intensive computing power they are much faster than the CPU.

1.3.2 CUDA

Compute unified device architecture is referred as CUDA. It has been introduced by NVIDIA in 2007. It can develop number of applications for GPU's that are highly parallel in nature and run on hundreds of GPU processor cores in parallel. CUDA builds threads which access fast shared memory and carry out parallel execution.

The CUDA programming language is much similar with C language and has a high learning curve. CUDA includes some API calls, various libraries in order to access the GPU specific features. It has some specific functions called kernels.

- CUDA Architecture

CUDA serves the efficient programming environment for number of cores of graphics processor to run in parallel and provides high computations. Applications that run on the CUDA architecture can take advantage of an installed base of over one hundred million CUDA-enabled GPUs in desktop and notebook computers, professional workstations, and supercomputer clusters. CUDA architecture comes with few basic components are listed below.

1. NVIDIA GPUs are designed for highly performance and parallel computation which best served by parallel compute engines.
2. OS kernel-level support for hardware initialization, configuration, etc.
3. User-mode driver, which provides a device-level API for developers
4. PTX instruction set architecture (ISA) for parallel computing kernels and functions

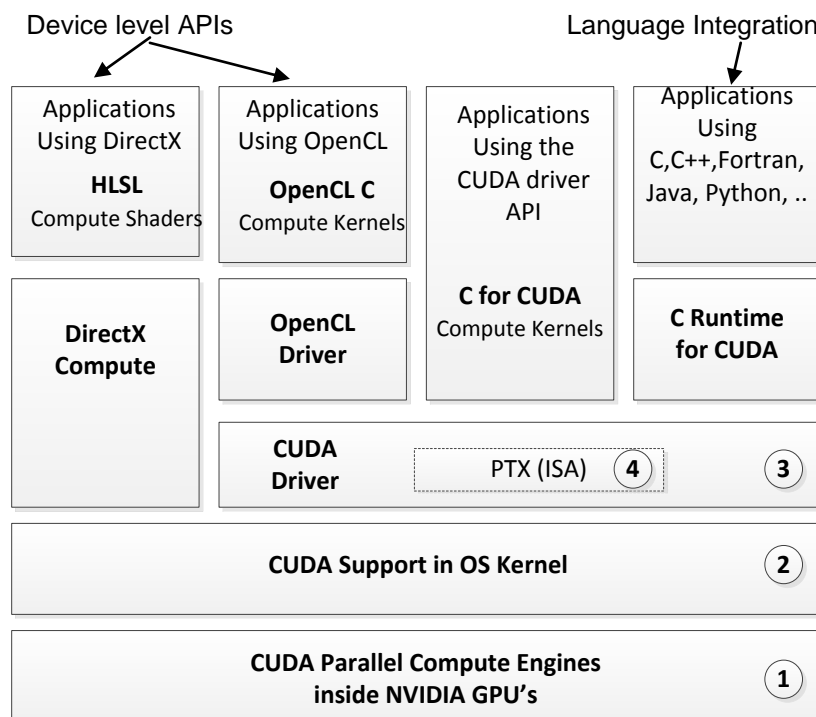


Figure 1.3 Basic CUDA Architecture

- CUDA programming

NVIDIA invented the parallel computing platform and programming model in terms of CUDA. It enables dramatic increase in computing performance by harnessing the power of the graphics processing unit (GPU). Using this technology the pattern matching algorithm for Genome sequence can be made optimized. NVIDIA developed the CUDA programming model and software environment to let programmers write scalable parallel programs using a straightforward extension of the C language. The CUDA programming model enables programmer to expose substantial fine-grained parallelism sufficient for utilizing massively multithreaded GPUs, while at the same time providing scalability across the broad spectrum of physical parallelism available in the range of GPU devices [9].

The Figure 1.4 shows CUDA programming structure. Grid represents the cores which process the data by partitioning it into smaller blocks. Each block can access maximum 512 threads (hardware dependent). At a time only one kernel is enabled.

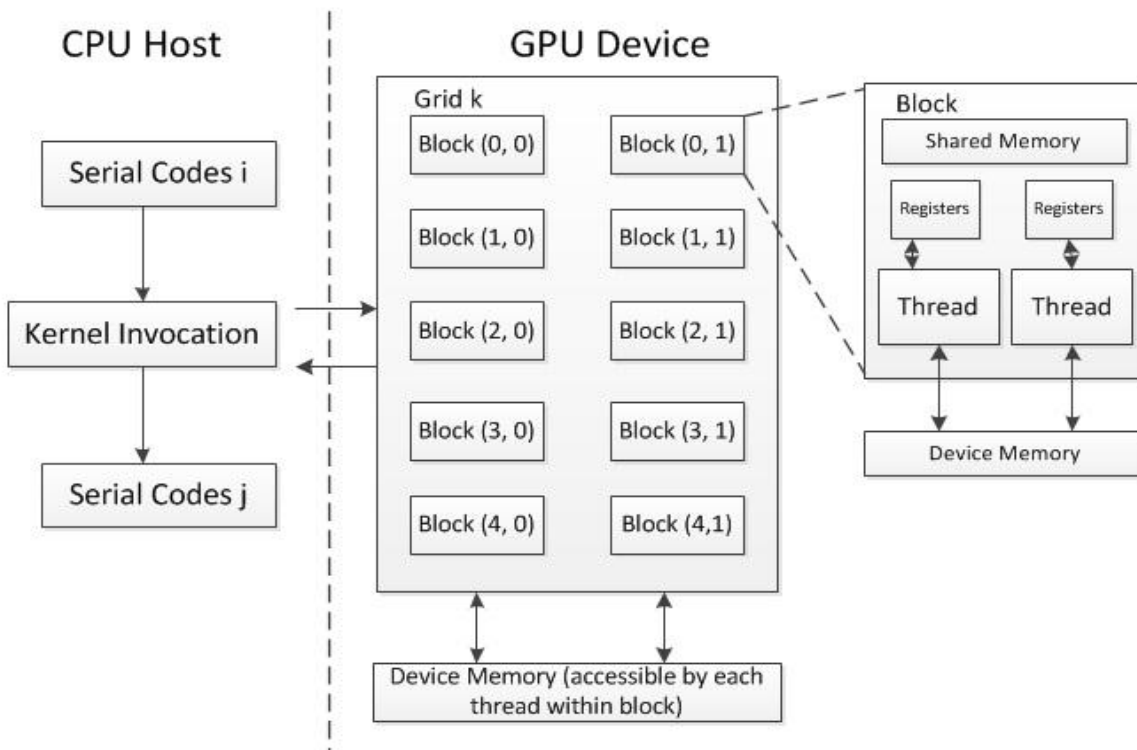


Figure 1.4 CUDA Programming Architecture

1.4 Project Objective

1.4.1 Objectives

The thesis discusses number of pattern matching algorithms and their performance differences. A simple string matching algorithm has chosen in order to apply the maximum level of parallelization with less system overhead. This approach is applied for the process of cancer diagnosis by matching several gene patterns in input sequence and draw inferences from result. Although existing cancer diagnosis technologies give effective results for DNA sequence, high computation, high end systems, and expensive equipment are needed for accomplishing the task. The overall cancer related molecular science, responsible genes, existing cancer detection technologies are studied and appropriate gene data is collected to fulfill the thesis requirements. The research has been carried till completion setting some of objectives. These are as follows.

- Analyze different serial algorithms for pattern matching and their performance ratios.
- Build scalable parallel brute force algorithm for DNA sequence with necessary modification.
- Measure the GPU and CPU performance differences in terms of processing time and generate resultant graph.
- Various cancer gene patterns are run according to thread per pattern and verified whether mutated or not.
- To increase the sensitivity of the filtering mechanism, the problem of pattern matching that errors are also taken care of.

Chapter 2

Literature Survey

2.1 Study of string matching algorithms

2.1.1 Traditional pattern matching algorithm

The literature describes various traditional pattern matching methodologies like Naïve Brute force, Boyer Moore, Knuth Morris Pratt and Dynamic algorithms and some newly introduced efficient algorithms like MSMPMA[2], IBKMPM[3], EPMSPP[4] and IAEMA[6] along with their performance issues when applied for gene analysis. Pattern matching is used in various processes. Like codon optimization which is carried out to enhance the efficiency of the DNA expression vectors used in DNA vaccination and gene therapy by increasing protein expression.

- Naive Brute force

It is one of the simplest algorithms having complexity $O(mn)$. In this, First character of pattern P (with length m) is aligned with first character of text T (with length n). Then scanning is done from left to right. As shifting is done at each step it gives less efficiency.

- Boyer-Moore Algorithm [BM1977]

It performs larger shift-increment whenever mismatch is detected. It differs from Naïve in the way of scanning. It scans the string from right to left; unlike Naive i.e. P is aligned with T such that last character of P will be matched to first character of T . If character is matched then pointer is shifted to left to very rest of the characters of the pattern. If a mismatch is detected at say character c , in T which is not in P , then P is shifted right to m positions and P is aligned to the next character after c . If c is part of P , then P is shifted right so that c is aligned with the right most occurrence of c in P . The worst complexity is still $O(m+n)$.

- Knuth-Morris-Pratt [KMP1977]

This algorithm is based on automaton theory. Firstly a finite state automata model M is being created for the given pattern P . The input string T with $\Sigma = \{A, C, T, G\}$ is processed through the model. If pattern is present in text, the text is accepted otherwise rejected. The following (e.g. Fig. 1) is a NFA model, created for ACATG string pattern.

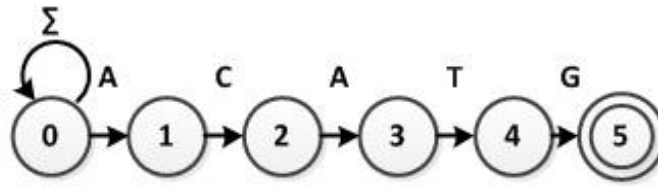


Figure 2.1 NFA for ACATG

For more simplification NFA is modified to DFA. The following (e.g. Fig. 2) is the equivalent DFA of the above NFA.

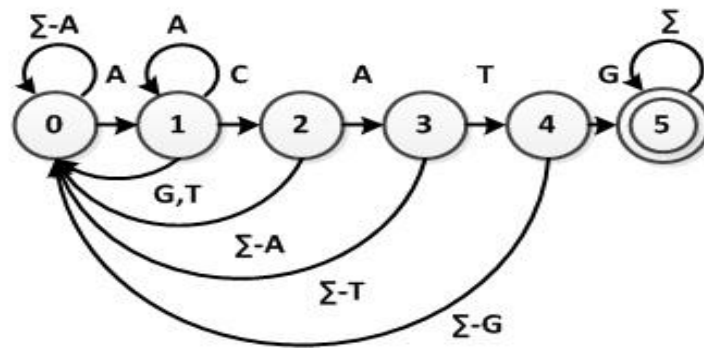


Figure 2.2 Deterministic Finite Automata for ACATG

But the only disadvantage of the KMP algorithm is that it doesn't tell the number of occurrences of the pattern.

Dynamic programming is the oldest and mostly used algorithm. Basically Needleman Wunsch and Smith waterman algorithm come under this approach. These are much more complex than the exact pattern matching. It involved solving successive recurrence relations recursively i.e smaller problems are solved in succession to solve the main problem. It has flexibility in adapting to different edit distance functions. That influences it's worth [10].

2.1.1 Enhanced matching Algorithms

- MSMPMA [10]

It is a simple one that carries multi pattern match with reduced number of comparison. This algorithm fixes index position and then compares the substrings of text to that of pattern until a match is found. Later generates the skip value on the basis of match number, this is taken further till the position reaches to $n-m$ where n being the text length and m is the pattern length. The skipping approach gives the number of occurrences of pattern with comparatively less number of shifts than other commonly used multiple pattern matching algorithms like Naïve brute force and Trie-matching. Hence it is faster. This algorithm has been implemented in a processor module to speed up the matching process with software implementation [5].

- IKPMPM [12]

This came up with better indexing technique. Here an index table is built to reduce the number of comparisons, And later partitioning the string and pattern (with some fix value k). First character of all partitions in string is compared to the first character of partitions of pattern. If a match is found it is processed, till the end of partition unless searching is stopped whenever mismatch occurs. This gives good performance for DNA related sequence application.

- EPMSPP [13]

It proposes even more efficient pattern matching approach called exact multiple pattern matching algorithms using DNA sequence and pattern pair. Instead of indexing each character of text, all 16 pairs of bases (A, C, T, and G) are indexed. This simplifies indexing and finds the pattern match on basis of pair indexing. Frequency of each pair of characters is found to get the number of occurrences of the pattern. This algorithm shows experimental results for a input sequence of 1024K. As the size of the data grows rapidly, especially in case of biological science where almost 3 billion bases length of DNA sequence is to be processed, it necessitates being in existence of such an algorithm which could handle the long length data efficiently.

Again IKPMPM[23] is brought with new improvement in which string matching algorithm is carried out with multithreading which speeded up the matching process by running multiple threads simultaneously where each thread applies matching process on a part of string.

Hence the term parallel computing came into existence. This was first invented by NVIDIA which provides the platform for parallel programming model by splitting the input sequence into streams and running all the streams in parallel. This [9] paper explains both string patterns and regular expression by compiling the patterns into a Deterministic finite automata. A sequence of n bytes can be processed using $O(n)$ operations irrespectively of speed. This is achieved because every state generates at most one new state. So the overall idea is aimed to use thread level parallelism for the DFA based matching process by splitting the input data strings into different chunks. Each chunk is processed in single thread independently which carries the same function for rest of the threads for the corresponding chunk.

Here the DFA works as in a string starting from initial state to the next state till it finds the matching character and then ends at final state represented by 2D table of states versus characters. This DMA model is brought out some of the ideas of few tools that use regular expression like `grep(1)`, `flex(1)` and `pcre(3)`. Which later convert the regular expression into DFA. In this literature they have used Thompson algorithm to convert from regular expression to NFA then subset construction algorithm is used to convert into DFA model. The basic idea of subset construction is to define a DFA in which each state is a set of states of the corresponding NFA. This processing achieves $O(1)$ processing cost for each scanned character of the input[4].

- IAEMA [14]

This article defines an Intelligent and efficient pattern matching algorithm which reduces number of comparisons through checking operation. Checking operation increases the skip distance. The effective technique is used here in two phases. In first phase an appropriate candidate key from the text string is searched. And another phase verifies the key to that of pattern in detail. It is an improvement over ECSA [12] algorithm.

A string matching algorithm is a succession of checking and skipping, where the aim of a good algorithm is to minimize the work done during each checking and to maximize the length distance during the skipping. An algorithm can skip to the next position in the text without missing any pattern occurrence. Most of the string matching algorithms pre-process the pattern before the search phase to help the algorithm to maximize the length of the skips. The pre-processing phase in this new MRCA4m algorithm helps in increasing the performance of maximizing the length of the skips [6].

2.2 Pros and cons of matching techniques

Various pattern matching algorithms that have been used in different applications are surveyed and studied. The survey has been done on the different techniques used to carry out matching process. Each came with new improvement. The different techniques work differently with different processing time even though some reflect same complexity, brings out different results. The table (TABLE II) summarizes all the techniques studied along with their pros and cons.

ALGORITHMS	PROS	CONS
Naive Brute Force	Simple	More number of shifts
Boyer Moore	Reduced number of shifts	Bad Character shift
Knuth Morris Pratt	Efficient for single pattern match	Doesn't offer time advantage over Boyer Moore for exact pattern match.
MSMPMA	Skips number of comparisons Suitable for single pattern match	As the size of text increases complexity increases
IKPMPM	Efficient indexing method with k partitions Comparisons per character ratio are less than existing approaches.(< 0.6)	Doesn't work for approximate pattern match.
EPMSPP	Pair indexing improves indexing	Performance degrades with error inputs
IAEMA	Checking and skipping reduces comparisons Increases system performance	Decreases system performance with decrease in pattern length
Parallel Algorithm	Parallel computing Least Processing time	System dependent

Table 2.1 Algorithms with their pros and cons

Chapter 3

Design model

All cancers arise as a result of changes that have occurred in the DNA sequence of the Genomes of cancer cells. Over the past quarter of a century much has been learnt about these mutations and the abnormal genes that operate in human cancers. It is possible to obtain the complete DNA sequence of large number of cancer genomes due to the vast improvement in the biological studies. These studies will provide us with a detailed and comprehensive perspective on how individual cancers have developed [13]. This knowledge came up with certain gene patterns whose occurrences in the specific DNA sequences can derive the cancer results.

3.1 Code Flow

At the very beginning of the CUDA code's execution, code is compiled just like other c code. Its primary execution takes place in CPU. As the execution started all non-kernel functions getting executed on CPU and the execution of kernel code is being transferred to GPU. This way we get parallel execution on CPU and GPU. Once the memory transfer between CPU to GPU is done, without any impediments the rest of the execution is carried well otherwise execution will be halted. Pattern matching gives out the search results for presence of specific gene in DNA sequence. Different gene sequences are taken from well-known databases and genome project. This can be applied in cancer diagnosis by matching several gene patterns in input sequence and draw inferences from result. The simplest brute force technique is used for the matching so as to cope up with complexity and to prevent possible overhead occurring due to parallelization. There will be two input files, one carrying gene code for each gene; another will be set of files carrying the information. To extract pattern match from a large sequence it takes more time. In order to reduce searching time matching is carried out parallel that reduces the search time with accurate retrieval.

The basic idea behind using the parallel approach for cancer diagnosis is quite simple. The genes that are responsible for particular type of cancer are being organized (data is collected from well-known database NCBI and other genome projects). DNA is declared to be cancer prone unless each is gene pattern is exactly present in given human DNA.

The overall working strategy is explained with the Figure 3.1.

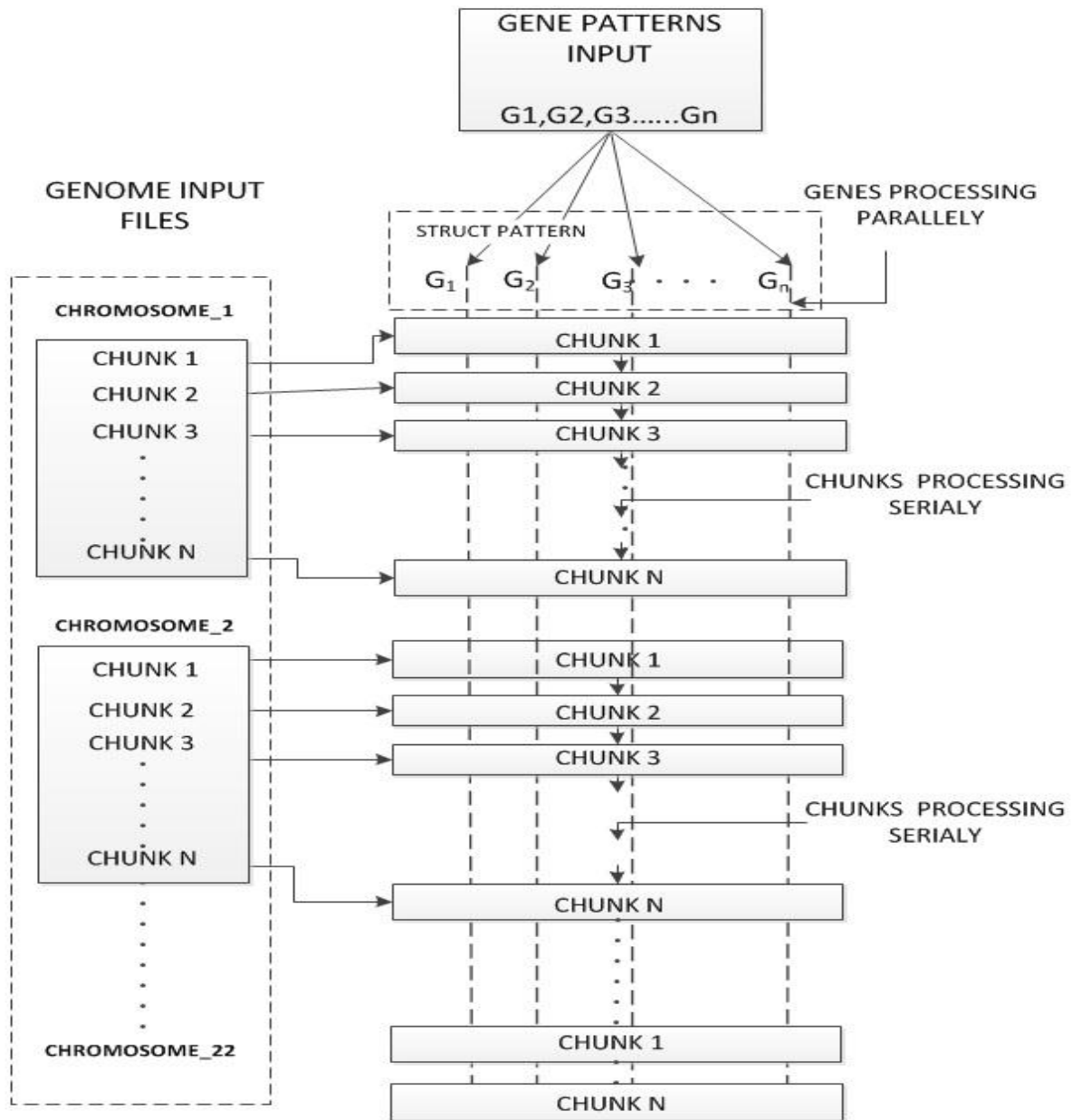


Figure 3.1 Code flow of the Pattern matching

Chapter 4

Implementation and Results

4.1 System requirements

The GPU used in this research work is TESLA C2070 on a HP Z420 workstation i7 core having 16 GB RAM operated on a 64 bit windows system. Latest release CUDA 5 is being configured with visual studio 2008 as platform. It is the NVIDIA computing processor designed to redefine high performance computing. It is based on the next –generation CUDA “Fermi” architecture. Compared to latest quad-core CPU’s Tesla C2070 computing processor deliver equivalent supercomputing performance at the rate 1/10th the cost and 1/20th the power consumption. It is built with 448 cores which delivers 515 Gflops double precision floating point performance and maximizes bandwidth between the host systems and tesla processors. It enables Tesla systems to work with virtually any PCIe-compliant host system with open PCIe x 16 slot. It can access maximum 1024 threads per block with 8 resident blocks per streaming processor. It possesses 6 GB of dedicate memory giving 2 compute capability which maximizes performance and reduces data transfers by keeping larger data sets in local memory that is attached directly to the GPU. ECC memory provides robustness in critical requirements for computing accuracy and reliability for workstations. Offer protection of data in memory to enhance data integrity and reliability for applications. Register files, L1/L2 caches, shared memory and DRAM all are ECC protected.

CUDA programming model is used in this implementation work. CUDA makes the computing engines of graphics processor units accessible to general purpose software developers through a standard programming language C, with an API to explicit the architecture parallelism. GPU allows thousands of threads to run in parallel.

4.2 System Results

Each pattern representing as a gene runs with a thread. As threads run in parallel, all the gene patterns are simultaneously searched in the chunks of text file. This way detection mechanism is achieved with less processing time.

4.2.1 Time versus Number of Patterns

The graph shows difference between the processing times of the serial and parallel implementation. As the number of patterns goes on increasing the processing time requires for the serial code also enhances. Against to that of parallel code, where processing almost same for all the patterns sets. In fact, it shows a better performance for more number of patterns. It would have been difficult to measure the performance difference for varying sizes of actual gene patterns so sample patterns are taken in order to maintain the accuracy of system performance. The results has been made for set of files with varying sizes ranging from 20-80 MBs and each is divided into fixed chunk size of 65535.

Table below shows the processing time differences between the implementation results achieved by parallel and serial code.

No of Patterns	Parallel	Serial
2	1.7	18.99
4	1.69	30.53
6	1.7	41.15
8	1.73	50.93

Table4.1 processing times for parallel and serial implementation with increasing patterns

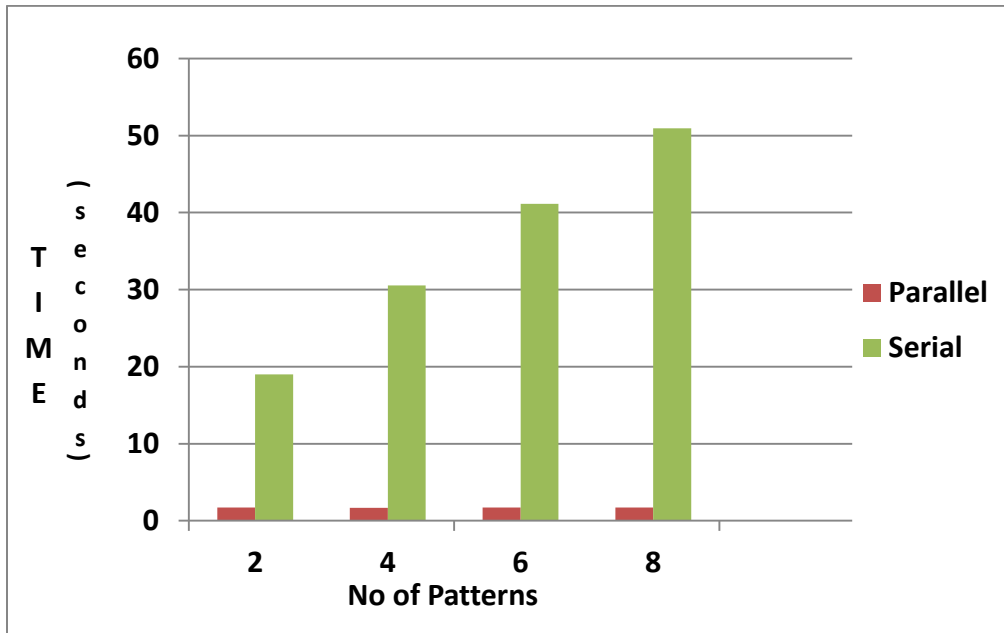


Figure 4.1 Time versus No of Patterns (with reference to Table 4.1)

The graph shows a vast variation in the serial performance but in case of parallel implementation the variation seems to be negligible. The fig 4.2 will highlight the minute differences for the parallel code with different sets of pattern.

No of Patterns	Processing Time
2	1.7
4	1.69
6	1.7
8	1.73

Table 4.2 processing times for parallel implementation with increasing patterns

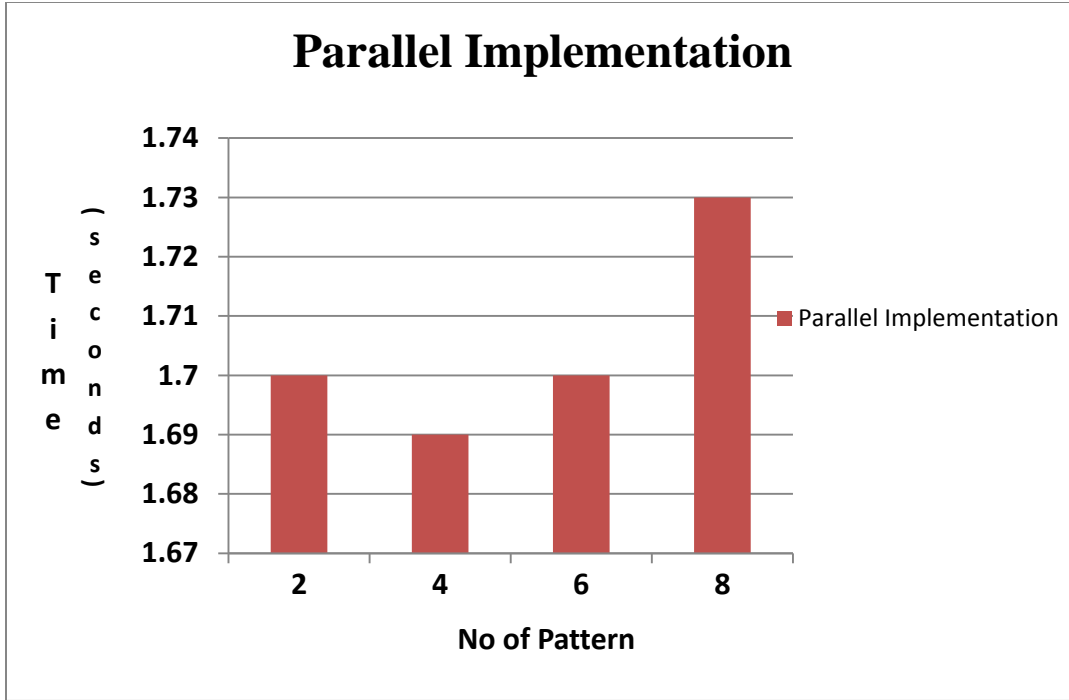


Figure 4.2 Time versus No of Patterns (with reference to Table 4.2)

4.2.2 Time versus Chunk size

In this project whole genome sequence is distributed into the different files as per the chromosome numbers. Each chromosome files with varying sizes. Each file is again divided into chunks so as to carry out the searching process efficiently. Fig is the graph of time versus chunk size. If the chunk size is increased, the processing time decreases in both the cases. But for parallel implementation the variation is negligible and serial implementation shows varies greatly. The parallel code gives 30 times more efficient than serial. The results has been made for set of files with varying sizes ranging from 20-80 MBs processing 8 number of patterns.

Chunk Size	Parallel	Serial
25000	1.93	17.31
35000	1.79	24.18
45000	1.72	67.6
55000	1.71	52.4
65000	1.69	46.24

Table 4.3 processing times for parallel and serial implementation with increasing chunk size

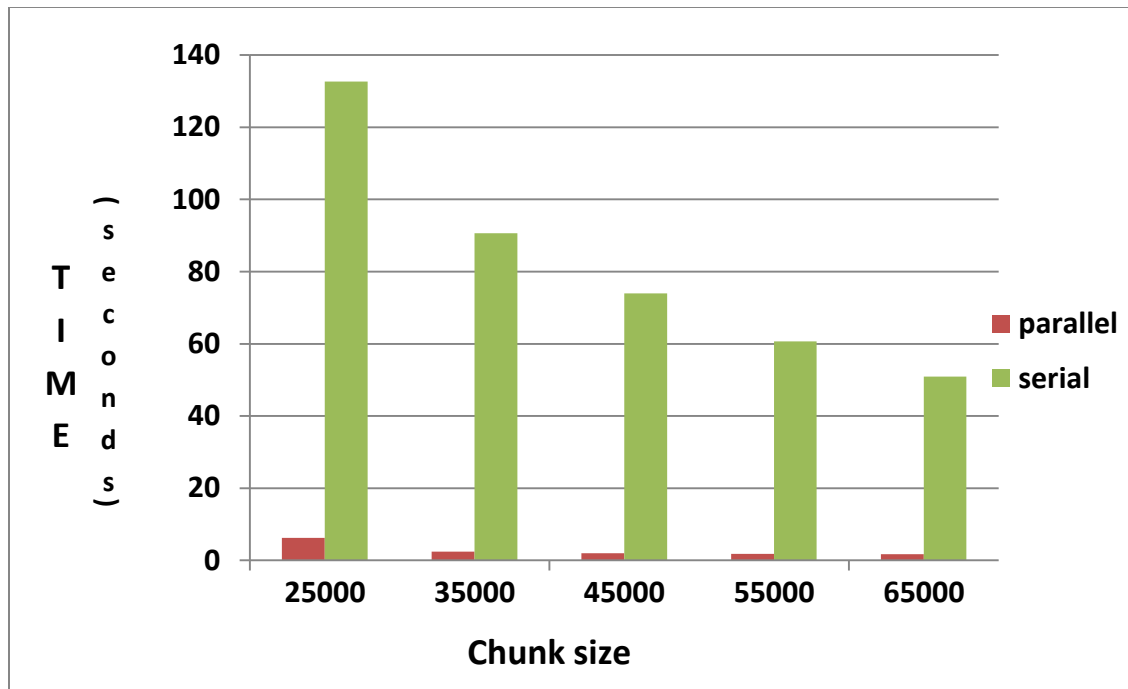
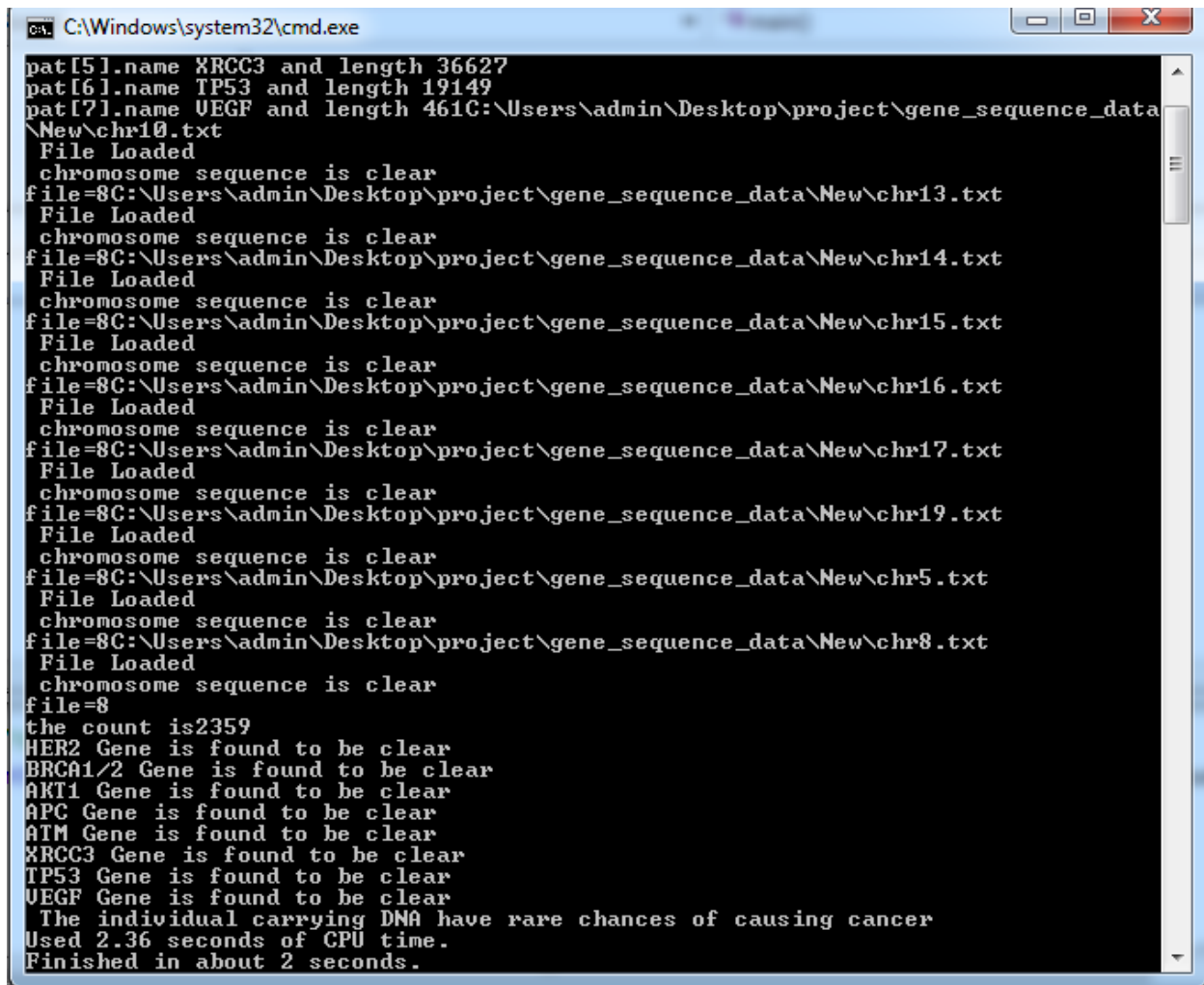


Figure 4.2 Time versus chunk size (with reference to Table 4.3)



```
C:\Windows\system32\cmd.exe
pat[5].name XRCC3 and length 36627
pat[6].name TP53 and length 19149
pat[7].name VEGF and length 461C:\Users\admin\Desktop\project\gene_sequence_data
New\chr10.txt
File Loaded
chromosome sequence is clear
file=8C:\Users\admin\Desktop\project\gene_sequence_data\New\chr13.txt
File Loaded
chromosome sequence is clear
file=8C:\Users\admin\Desktop\project\gene_sequence_data\New\chr14.txt
File Loaded
chromosome sequence is clear
file=8C:\Users\admin\Desktop\project\gene_sequence_data\New\chr15.txt
File Loaded
chromosome sequence is clear
file=8C:\Users\admin\Desktop\project\gene_sequence_data\New\chr16.txt
File Loaded
chromosome sequence is clear
file=8C:\Users\admin\Desktop\project\gene_sequence_data\New\chr17.txt
File Loaded
chromosome sequence is clear
file=8C:\Users\admin\Desktop\project\gene_sequence_data\New\chr19.txt
File Loaded
chromosome sequence is clear
file=8C:\Users\admin\Desktop\project\gene_sequence_data\New\chr5.txt
File Loaded
chromosome sequence is clear
file=8C:\Users\admin\Desktop\project\gene_sequence_data\New\chr8.txt
File Loaded
chromosome sequence is clear
file=8
the count is2359
HER2 Gene is found to be clear
BRCA1/2 Gene is found to be clear
AKT1 Gene is found to be clear
APC Gene is found to be clear
ATM Gene is found to be clear
XRCC3 Gene is found to be clear
TP53 Gene is found to be clear
VEGF Gene is found to be clear
The individual carrying DNA have rare chances of causing cancer
Used 2.36 seconds of CPU time.
Finished in about 2 seconds.
```

Figure 4.3 Resultant output of the pattern matching algorithm for DNA sequence

Above fig(Figure 4.3) shows the output of the execution of the parallel implementation of pattern matching algorithm which diagnose the input DNA sequence by detecting exact presence of provided genes. The specified output first displays the name of all the input genes with corresponding lengths. Then each pattern is simultaneously searched in the several files loaded into directory name new. As the input DNA sequence belongs to normal (not having cancer susceptibility) individual, it generates true negative diagnosis.

Chapter 5

Conclusion and Future Scope

5.1 Application

Cancer is one of the severe diseases causing one in eight deaths worldwide. It can be cured if detected at the very first stage where the cancer cells stay fixed in their area. In stage two it will start to spread. When it spread to muscles enters in third stage. It may cause organ failure. Last stage is the deadliest and inescapable. Success rate of recovery is highest if cancer is detected in the first stage. Early stage cancer detection can be achieved through genetic tests. The research stated in this paper can be used in Gene testing technology to recognize gene differences by analyzing DNA sequence rather to use complex, expensive equipment provided accurate molecular data is available.

5.2 Conclusion

The report investigates an efficient and simple mechanism for cancer detection. From the obtained results, an individual is verified whether he/she has chances of getting cancer in future or not though his/her DNA. An ordinary middle class individual may find it prohibitive to use existing diagnosis technology as it is bit expensive. The research is done on GPU using CUDA programming model, accelerating the searching process. This has led significant improvement over serial analysis as it is implemented on GPU.

5.3 Future Scope

There are around 200 cancer types worldwide that can affect human body. The work has been done to diagnose commonly occurring cancer types. In future if molecular science has brought up new study results about cancer genes for other types of cancer, the work can be extended for all types of cancers.

Chapter 6

Publication

The thesis has been made along with research survey paper titled “Survey Paper on the Pattern Matching Algorithm for DNA sequences” has been published in the conference:

International Conference on Recent Trends in Computer Science and Engineering, ISBN; 978-93-81583-89-05, 24th April, 2013, Bangalore.

Bibliography

- [1] Michael Garland, Scott Le Grand, John Nickolls from NVIDIA; Joshua Anderson, Iowa State University and Ames Laboratory , Jim Hardwick Techni Scan Medical Systems “Parallel computing experiences with CUDA”, IEEE. 0272-1732/08, 2008.
- [2] Chung W. Ng, Bio Chem, “Inexact Pattern Matching Algorithms via Automata” 218, March 19, 2007.
- [3] Flexible Pattern Matching in Strings: Practical On-Line Search Algorithms for Texts and Biological Sequences by Gonzalo Navarro and Mathieu Raffinot .
- [4] Charalampos S. Kouzinopoulos and Konstantinos G. Margaritis “Parallel and Distributed Processing Laboratory String Matching on a multicore GPU using CUDA”. 13th Panhellenic Conference on Informatics 2009.
- [5] Genome sequencing <http://www.genome.gov/>
- [6] Boyer Moore Algorithm <http://orion.lcg.ufrj.br/Dr.Dobbs/books/book5/chap10.htm>
- [7] Genetics home reference <http://ghr.nlm.nih.gov/handbook/basics/dna> DNA concept
- [8] GPU information <http://www.nvidia.in/object/gpu-computing-in.html>
- [9] Shrenik Shah, Harvard University, APPLIED MATHEMATICS CORNER “DNA Computation and Algorithm Design” Cambridge, MA 02138, 2009.
- [10] Ziad A.A Alqadi, Musbah Aqel & Ibrahiem M.M.EI Emary, “Multiple Skip Multiple Pattern Matching algorithms”, IAENG International. Vol 34(2), 2007.
- [11] M Yazid M Saman, M Nordin A Rahman, Aziz Ahmad, “A Minimum Cost Process in Searching for a Set of Similar DNA Sequences” May 27-29, 2006 (pp348-353), International conference-2006.
- [12] Raju Bhukya, DVLN Somaya julu, “An Index Based K Partition Multiple Pattern Matching Algorithm” Proc. Of International Conference on Advances in Computer Science 2010 pp 83-87.
- [13] Raju Bhukya, DVLN Somaya julu, “Exact Multiple Pattern Matching Algorithm using DNA Sequence and Pattern Pair” International Journal of Computer Applications (0975 – 8887) Volume 17– No.8, March 2011.

- [14] Mahmoud Moh'd Mhashi, "An Intelligent and Efficient Matching Algorithm to Finding DNA Pattern", IMACST: VOLUME 3 NUMBER 1 FEBRUARY 2012.
- [15] Alexander Gee Research into GPU accelerated pattern matching for applications in computer security, Department of Computer Science and Software Engineering University of Canterbury, Christchurch, New Zealand. November 4, 2009.
- [16] Peter Boyle and Bernard Levin "World cancer report 2008" World Health organization, International agency for research on Cancer Research UK August 2012.
- [17] International Agency for Research on Cancer Press Release N 210 "Indian Cancer statistic, a model to be followed" 28 march 2012.
- [18] American Cancer Society, "Cancer Facts & Figures 2013", Atlanta: American Cancer Society; 2013.
- [19] Genes responsible for Leukemia http://www.bioinformatics.org/legend/leuk_db.htm
- [20] Simon A. Forbes, Gurpreet Tang, Nidhi Bindal, Sally Bamford, Elisabeth Dawson, Charlotte Cole, Chai Yin Kok, Mingming Jia, Rebecca Ewing, Andrew Menzies, Jon W. Teague, Michael R. Stratton and P. Andrew Futreal COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer D652–D657 Nucleic Acids Research, 2010, Vol. 38, Database issue Published online 11 November 2009 doi:10.1093/nar/gkp995.
- [21] Michael R. Stratton, Peter J. Campbell, & P. Andrew Futreal "The Cancer Genome", Vol 458, 9 April 2009, doi:10.1038: nature07943.
- [22] Sinha R, Anderson DE, McDonalds SS, Greenwald P "Cancer Risk and Diet in India", Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health.
- [23] S.Nirmala Devi, S. P. Rajagopalan "An Index Based Pattern Matching using Multithreading", International Journal of Computer Applications (0975-8887) Volume 50- No. 6, July 2012.