Various String Matching Algorithms for DNA Sequences to Detect Breast Cancer using CUDA Processors

Nagaveni V
Department of Computer Science and Engineering
Bharathiar University
Coimbatore, 641046,Tamilnadu, India
nagaveni@acharya.ac.in

G T Raju
Department of Computer Science and Engineering
RNS Institute of Technology
Bangalore -560061, Karnataka, India
gtraju1990@yahoo.com

***Abstract*— The main aim of string matching algorithm is to locate the appearance of a specific pattern in an array of larger size text. String matching algorithms has been used in many applications such as DNA analysis. This report introduces a new 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; many new and efficient algorithms are required to carry out the cancer diagnosis. This can be attained easily by applying parallelization techniques using GPU using CUDA programming model.**

***Index Term*— Parallel Programming; CUDA Programming; String matching algorithms; DNA sequencing**

I. INTRODUCTION

Deoxyribonucleic acid, more commonly known as *DNA*, is a complex molecule that contains all of the information necessary to build and maintain an organism. 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. Structure of DNA is as shown in figure 1.*Genes* are segments of DNA located on chromosomes. *Genes* contain the codes for the production of specific proteins. 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. This paper 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.

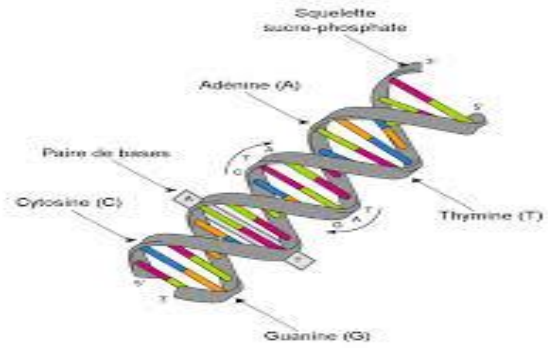

Fig. 1. Structure of DNA

II. RELATED WORK

Numerous works have been done till now that analyze and compare parallel programming paradigms for multicore clusters. To mention a few, there are different techniques to detect different types of cancers and to cure them. In that females are facing now a day's breast cancer problems from which many deaths are happening. Hence GPUs are used to identify at earlier stage and cure them. But all these results vary depending on the type of the problem solved, the algorithms used and the features of the hardware architectures used ,which makes research in this area even more significant.

III. DNA SEQUENCING

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 the process of determining the DNA *sequence* may come in useful in practically any *biological* research.DNA sequencing is used to determine the sequence of individual genes, larger genetic regions, full chromosomes or entire genomes.

A. 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.

Cancer is responsible for majority of deaths worldwide. There are more than hundred types of cancer diseases which originate from most of the cell types and organs of the human body. All cancers occur due to mutation in the gene sequence.

The approximate data disclosed by research centers like International Agency for research on cancer (IARC) addresses that India has lower cancer rates than foreign countries as shown in table 1.1.

Table 1.1
Table 1.1 is the estimated percentage of cancer amongst male and female

Cancer	Men	Cancer	Women
Lung & Bronchus	28%	Lung & Bronchus	26%
Prostate	10%	Breast	14%
Kidney & renal pelvis	3%	Brain/other nervous system	2%

B. 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. The most common cancer diagnostic methods are Biopsy, Endoscopy, Diagnostic imaging, Blood test, Pap test and genetic test. One of the important cancers detecting method is a Genetic testing. In simple words, Genetic testing is a reading a DNA code to identify abnormalities.

1) Genetic Testing

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 is categorized 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.

2) 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 this happens, the cell grows out of control, which can lead to cancer. E.g.RET.

Tumor Suppresser 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.

C. BRCA1 and BRCA2: BRCA1 and BRCA2

are human genes that produce tumor suppressor proteins. These proteins help repair damaged DNA and, therefore, play a role in ensuring the stability of the cell's genetic material.

Specific inherited mutations in BRCA1 and BRCA2 increa se the risk of female breast and ovarian cancers, and they have been associated with increased risks of several additional types of cancer.

The gene responsible for breast cancer is listed in the Table 1.2.

Table 1.2
Cancers with their responsible genes.

Cancer	Responsible Gene	Gene Location	Chromosome sequence
Breast	HER2,BRCA1/2,AKT 1	17q12,17q21, 14q32	Chromosome: 17 Chromosome: 17, Chromosome: 14

1) Breast cancer detection and diagnosis

Today, medical imaging systems using NVIDIA GPUs give doctors and patients more accurate results and faster diagnoses than have been possible with traditional scanning systems. As a result, our technologies are helping to improve breast cancer detection to find cancer earlier, thus improving patient survival rates and profoundly and positively affecting women's health and wellbeing. Several options are available for managing cancer risk in individuals who have a known harmful *BRCA1* or *BRCA2* mutation. These include enhanced screening, prophylactic (risk-reducing) surgery, and chemoprevention.

- **Enhanced Screening.** Some women who test positive for *BRCA1* and *BRCA2* mutations may choose to start screening at younger ages than the general population or have more frequent screening. For example, some experts recommend that women who carry a harmful *BRCA1* or*BRCA2* mutation undergo clinical breast examinations beginning at age 25 to 35 years (17). And some expert groups recommend that women who carry such a mutation have a mammogram every year, beginning at age 25 to 35 years.
- **Prophylactic (Risk-reducing) Surgery.** Prophylactic surgery involves removing as much of the "at-risk" tissue as possible. Women may choose to have both breasts

removed (bilateral prophylactic mastectomy) to reduce their risk of breast cancer. Surgery to remove a woman's ovaries and fallopian tubes (bilateral prophylactic salpingo-oophorectomy) can help reduce her risk of ovarian cancer. Removing the ovaries also reduces the risk of breast cancer in premenopausal women by eliminating a source of hormones that can fuel the growth of some types of breast cancer.

- **Chemoprevention.** Chemoprevention is the use of drugs, vitamins, or other agents to try to reduce the risk of, or delay the recurrence of, cancer. Although two chemo preventive drugs (tamoxifen and raloxifene) have been approved by the U.S. Food and Drug Administration(FDA) to reduce the risk of breast cancer in women at increased risk, the role of these drugs in women with harmful BRCA1 or BRCA2 mutations is not yet clear.

D. *Parallel Programming Model*

1) Graphics Processing Unit

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.

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. Comparison study of CPU and GPU is shown in Figure 2 for different application areas.

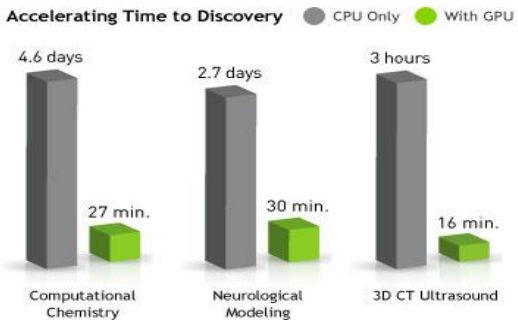

Fig. 2.

IV. Comparison Study of CPU and GPU

Compute unified device architecture is referred as CUDA 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.

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.

A. *CUDA programming*

NVIDA 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.

This paper 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.

B. Analyze different serial algorithms for pattern matching and their performance ratios.

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

a) 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.

b) 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. The worst complexity is still O(m+n).

c) 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 $\sum = \{A, C, T, G\}$ is processed through the model. If pattern is present in text, the text is accepted otherwise rejected. But the only disadvantage of the KMP algorithm is that it doesn't tell the number of occurrences of the pattern.

C. *Enhanced matching Algorithms*

*1) MSMPMA (*Multiple Skip Multiple Pattern Matching Algorithm)

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.

2) IKPMPM (Index based K partition Multiple Pattern Matching algorithm*)*

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). This gives good performance for DNA related sequence application.

3) EPMSPP (Exact Multiple Pattern Matching Algorithm)

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. This algorithm shows experimental results for a input sequence of 1024K.

The table 4.1 summarizes all the techniques studied along with their pros and cons.

Table4.1
Algorithms with their pros and cons

ALGORITHMS		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
Parallel Algorithm	Parallel computing Least Processing time	System dependent

4) 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.

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.

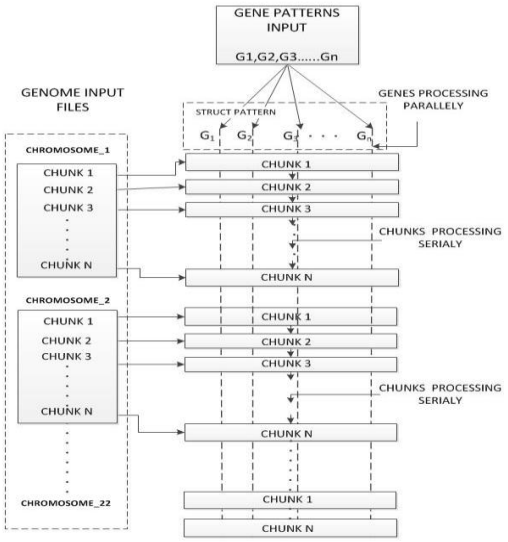

Fig. 3. Code flow of the Pattern matching

V. Implementation and Results

A. *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.

B. *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.

1) *Time versus Number of Patterns*

The graph of figure 4 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.

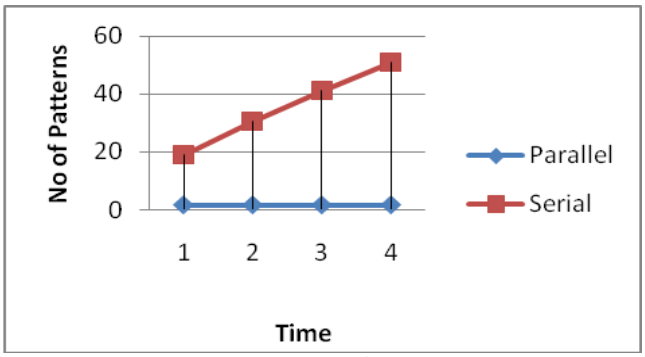

Fig. 4. Time versus No of patterns graph

The graph shows a vast variation in the serial performance but in case of parallel implementation the variation seems to be negligible.

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. Figure 5 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 have been made for set of files with varying sizes ranging from 20-80 MBs processing 8 numbers of patterns.

Table 5.1
processing times for parallel and serial implementation with increasing chunk size

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

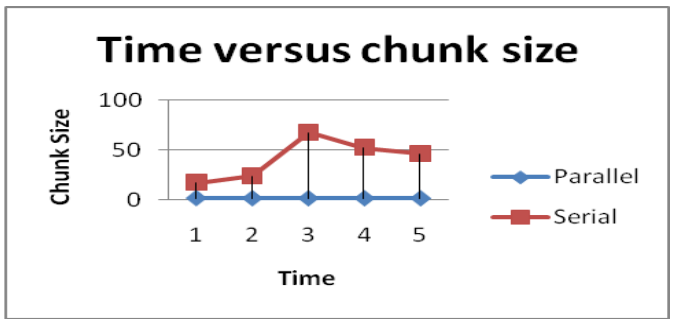

Fig. 5. Time versus chunk size (with reference to Table 5.1)

Above fig (Table 5.1) shows the output of the execution of the both parallel and serial implementation of pattern matching algorithm which diagnose the input DNA sequence by detecting exact presence of provided genes.

VI. . CONCLUSION

The report investigates an efficient and simple mechanism for breast cancer detection. From the obtained results, an individual is verified whether he/she has chances of getting breast 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. Availability of new techniques for treatment are also discussed. This has led significant improvement over serial analysis as it is implemented on GPU. Future lines of work include the development of other solutions for new symptom types of BRAC1 and BRAC2 .

REFERENCES

[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] Genetics home reference http://ghr.nlm.nih.gov/handbook/basics/dnaDNA concept

[6] GPU information http://www.nvidia.in/object/gpu-computing-in.html

[7] Shrenik Shah, Harvard University, APPLIED MATHEMATICS CORNER "DNA Computation and Algorithm Design" Cambridge, MA 02138, 2009.

[8] Ziad A.A Alqadi, Musbah Aqel & Ibrahiem M.M.EI Emary, "Multiple Skip Multiple Pattern Matching algorithms", IAENG International. Vol 34(2), 2007.

[9] 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.

[10] 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.

[11] Mahmoud Moh'd Mhashi, "An Intelligent and Efficient Matching Algorithm to Finding DNA Pattern", IMACST: VOLUME 3 NUMBER 1 FEBRUARY 2012.

[12] 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.

[13] Peter Boyle and Bernard Levin "World cancer report 2008" World Health organization, International agency or research on Cancer Research UK August 2012.

[14] International Agency for Research on Cancer Press Release N 210 "Indian Cancer statistic, a model to be followed" 28 march 2012.

[15] Michael R. Stratton, Peter J. Campbell, & P. Andrew Futreal "The Cancer Genome", Vol

[16] 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.