Abstract–The goal of this project is to apply high-performance pattern recognition and protein binding techniques to treat cancer. The objective of this proposal is to develop pilot data to write proposals to the external cancer research funding agencies. The research approach is as follows: (i) Generate digital/bit (0 or 1) images from picture images, such as magnetic resonance imaging (MRI), of the patient and/or patient's blood-relatives, (ii) help discover the most accurate diagnosis as fast as possible by applying high-performance pattern recognition on digital/bit images, and (iii) help treat cancer using the combined solutions of traditional practice and protein binding analysis. By year 2030, cancer is expected to become the number one cause of death in the United States, exceeding heart disease. Many U.S. government and other agencies including the National Cancer Institute (NCI) are providing funds to develop effective solutions for detection and treatment of cancer. The primary investigator (PI), an expert in high performance computing, has been collaborating with Dr. M.F. Islam (Oncology Specialist at the Univ. of Pittsburgh Medical Center), Dr. S. Anant (Associate Director of Cancer Prevention and Control, Univ. of Kansas Cancer Center), and Dr. M.J. Uddin (Research Assistant Professor of Biochemistry at Vanderbilt Univ. School of Medicine) to develop time- and cost-effective solutions for cancer treatment.

1. INTRODUCTION

According to a report from the American Society of Clinical Oncology, cancer will become the leading cause of death in the United States in the next 15 years. The number of new cancer cases is expected to be 2.3 million annually by 2030 [1]. *(References are available upon request.)* NCI at the National Institutes of Health (NIH) has increased its budget from $4.8 billion in year 2013 to $5.1 billion in year 2014 [2]. Many other organizations including American Cancer Society [3], American Society of Clinical Oncology [4], Breast Cancer Research Foundation [5], and Damon Runyon Cancer Research Foundation [6] are awarding research grants to develop better solutions to prevent and control cancer.

Based on our recent studies, discussions with oncology specialist, university researchers, and information technology (IT) experts, the growing demand for time- and cost-efficient solutions to determine and treat cancer is fueling the interest in developing high-performance pattern recognition and protein binding analysis techniques. In designing methods for high performance computing, both the extent of the difficulty and the nature of the proposed solutions depend radically on programming (sequential to parallel to multithreading) and architectural (single-core to multicore to manycore) constraints [7-9]. Multithreaded manycore computing is more natural and acceptable technique to improve performance (i.e., processing speed) [10, 11]. The success of fast detection and proper treatment of cancer depends on speedy/accurate analysis of a large number of images, conventional drugs, and protein binding results. Therefore, we propose to applying new high-performance pattern recognition and protein binding techniques for treating cancer.

2. BACKGROUND AND MOTIVATION

Pattern recognition is the basis for computer-aided diagnosis (CAD) systems that supports the physician's interpretations and findings [12-14]. Simply speaking, pattern recognition can be defined as dealing with feature extraction and classification [15]. Pattern recognition is a complex branch of machine learning (construction and study of algorithms). Image processing techniques (where the input may be a photographic image and the output may be a set of characteristics related to the image) are widely used for pattern recognition. Cancer detection using pattern recognition is very complex; it requires processing many images and becomes extremely time consuming.

All drugs used to treat cancer usually cause side effects and the side effects of each drug may vary from person to person [16-18]. Therefore, drugs should be tested for their tendency to bind with proteins using a protein binding evaluation. Protein binding defines the ability of proteins to form bonds with other substances such as drugs [19-20]. If a drug is 75% bound to a binding protein, then 25% of the drug is free (i.e., active) to cause pharmacological effects in the system. If drug A saturates a certain binding protein and then drug B is not able to bind to it, then there would be a higher concentration of

unbound B to affect the system. Analyzing the impact of protein bindings of existing drugs (for all kinds of proteins and drugs) requires enormous amount of time.

Scientists from many research institutions including the Univ. of California at Berkeley, Brown Univ., and Washington Univ. in St. Louis are excepting that computing will provide the best skills to treat cancer in the next decade [21-25]. Various computing tools, such as MATLAB, are developed to achieve high performance. MATLAB (widely used in academia and industry for simulating algorithms) has been updated into parallel MATLAB in the mid-2000s for better performance [26, 27]. Compute Unified Device Architecture (CUDA, developed in the mid-2000s) is a parallel programming platform for NVIDIA graphics processing unit (GPU) for high performance computing [28]. In our previous studies, we explore both MATLAB and CUDA for solving a weighted Jacobi iteration with a structured matrix that is derived from a finite element discretization [29-36]. The CUDA approach is proven to be the faster and efficient way of programming for high performance. Therefore, CUDA-accelerated high performance computing should be very effective for fast detection and proper treatment of cancer.

3. METHODOLOGY/APPROACH

3.1 Project Description: Applying pattern recognition and protein binding analysis to treat cancer

In this work, pattern recognition algorithms through image processing are applied to help find the most accurate diagnosis for any input MRI-type images by performing "most likely" matching of the inputs. Also, protein binding technique is utilized in addition to analyzing the conventional practices to help determine the exact dose of the correct drugs to treat cancer. The schematic diagram in Figure 1 illustrates the major steps and work-flow of the proposed methodology.

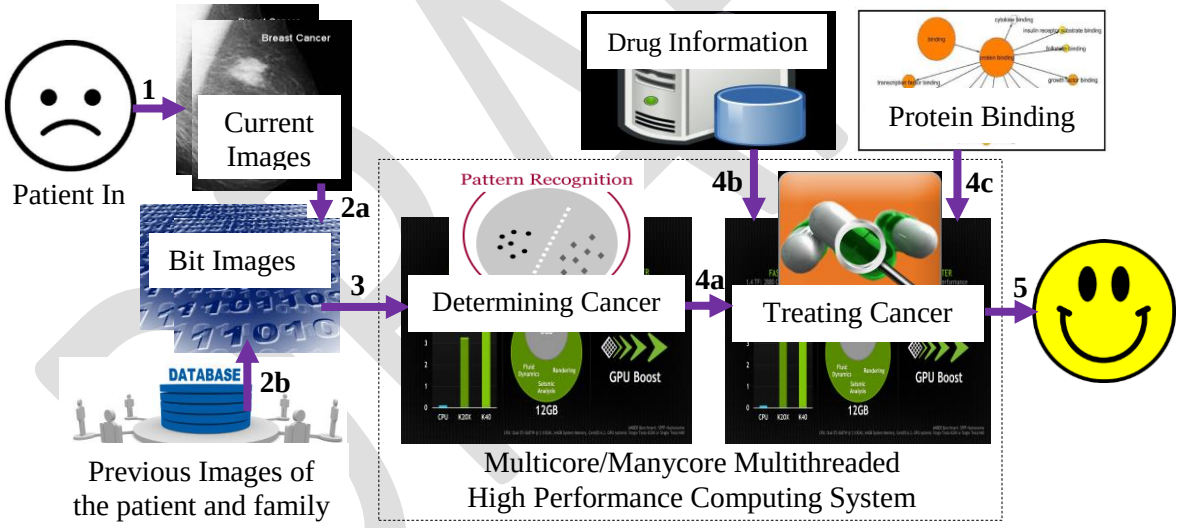

Figure 1: Schematic diagram and work-flow of the proposed methodology

First (new) MRI-type images are collected from the patient's body. Then (each of) the newly collected images and any previous images from the patient and/or patient's blood-related relatives are converted into digital/bit (0 or 1) images. All digital images are then processed on a state-of-the-art high performance computing system using pattern recognition techniques to decide any existence and/or severity of cancer. If cancer is detected, an initial list of drugs is selected. Finally, protein binding results are analyzed to finalize the drug(s) to treat the patient. Multicore multithreaded high performance computing system is used to speed up the pattern recognition and protein binding analysis.

3.1 Preliminary Work: CUDA Vs. MATLAB Implementation for Laplace's Equation [37]

In GPU computing, as illustrated in Figure 2, the initialization and serial portions are executed in the central processing unit (CPU). The CUDA/C application sends the data and code for the parallel portions to the GPU card. For each parallel part, multiple threads are generated. Generating and organizing the

threads (code and data) is the key for performance improvement. Many threads are executed in the GPU cores concurrently in parallel. After the executions are done, the results are sent back to the CPU from the GPU. This module only parallelizes the computationally expensive section of the problem. As a preliminary attempt, we run the simulation programs on our CUDA server (CPU: Xeon E5506, 8 cores, 2.13 GHz, 8GB DDR3 RAM; GPU: Telsa C2075, 448 cores, 6GB GDDR5 memory). Figure 3 illustrates the speedup due to CUDA/C shared memory and parallel MATLAB solutions over MATLAB (sequential) implementation of Laplace's equation (for electric charge distribution on a thin 2-D surface). According to the experimental results, the speedup due to CUDA/C starts increasing for large problem size (N > 1024). For N = 4096, CUDA/C implementation achieves more than 253x speedup. This work strongly suggests that CUDA accelerated GPU computing has potential to execute the large number of computations that are required for cancer detection and treatment.

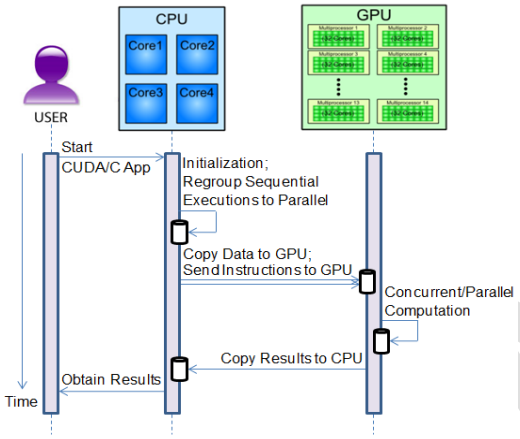

Figure 2: Data exchange mechanism.

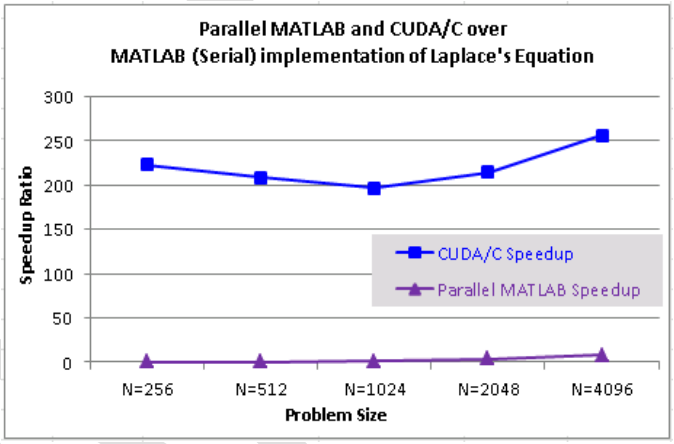

Figure 3: Speedup due to CUDA/C and MATLAB.

3.3 Justification of the Project

There are growing demands to developing better treatment for cancer. Major U.S. government funding agencies including NIH and many non-profit organizations including American Cancer Society are investing increasing amount of money on cancer research as it is a radically growing threat. Therefore, this is the right time for this research. Although writing high-performance multicore/multithreaded program is difficult [38, 39], various parallel programming languages, libraries, APIs, and models have been created [40-44]. Most contemporary personal computers (PCs) and Laptops come with built-in GPU card(s) [45-47]; the CPU-GPU combination provides an excellent hardware support for high performance computing. Therefore, the pattern recognition and protein binding analysis methodology, two outcomes of this project, can be easily integrated into any existing computing systems.

[Intentionally left incomplete]