A Multithreaded Model for Cancer Tissue Heterogeneity: An Application

Anik Chaudhuri, Shabnam Choudhury, Anwoy Kumar Mohanty, Manoranjan Satpathy

***Abstract*—Studying the heterogeneity in cancerous tissue is challenging in cancer research. It is vital to process the real-world data efficiently to understand the heterogeneous nature of cancer tissue. GPU compatible models, which can estimate the subpopulation of cancerous tissue, are fast if the size of input data, i.e., the number of qPCR (quantitative polymerase chain reaction) gene expression reading is extensive. In the real world, we rarely get that much data to reap the benefits of a GPU's parallelism. Real experimental data from fibroblasts are much less, and models using those data on a GPU are slower than the CPU multithreaded application. This paper will show a method to run GPU-compatible models for cancer tissue heterogeneity on a multithreaded CPU. Further, we also show that the model running on a multithreaded CPU is faster than the model running on a GPU with real experimental data.**

***Index Terms*—Bayesian methods,hierarchical model, Markov chain Monte Carlo, Metropolis-Hastings algorithm, Graphics processing unit, OpenMP**

I. Introduction

TUMOR cells are heterogeneous [1] and [2]. The clonal evolution models suggest that tumor cells accumulate mutation as it progresses. This stepwise accumulation results in various sub-population in a tumor cell and makes it heterogeneous. Another dominant theory called the stem cell suggests that only a small portion of the tumor cell are dominant [3], [4], [5] and [6]. These theories suggest that cancer tissues are heterogeneous and pose specific challenges in cancer treatment. The sub-population reacts differently to a given therapy. It may also happen that a particular combination of drugs works for a patient but may not work for another patient with a different combination of sub-population. It is essential to know the proportion of these sub-populations to make appropriate decisions about therapy. A mathematical model that can account for the heterogeneous behavior of cancer tissue can provide better insight into cancer treatment. Therefore, it is imperative to model the heterogeneity of cancer tissue mathematically.

Authors in [7] discussed a hierarchical model to analyze cancer tissue heterogeneity. The authors in [7], used a Boolean network collection to model cancer, and the weights of each of those networks represented the proportion of each heterogeneous sub-population. A hierarchical model was used to demonstrate the relationship between gene expression measurements and the unknown parameters. In [8], the authors presented a parallelizable model to analyze cancer tissue heterogeneity. Unfortunately, this parallelizable model, which is compatible with a GPU's SIMD (single instruction multiple data) architecture, does not perform well for a small dataset. The parallelizable model running on a GPU is suitable with large synthetic data but not for real experimental data because the amount of real experimental data is significantly less. So, a GPU is not an ideal choice for such cases. This paper describes a method to run parallelizable models on a multithreaded CPU. We shall show that a multithreaded CPU's run time is less than the GPU for real experimental data.

II. Motivation

Relations between proteins and DNA is responsible for cellular interaction [2] and [9]. Using gene regulatory networks is an excellent way to model cell behavior and develop better therapy. Gene regulatory networks have been modeled by differential equation [10], deterministic and probabilistic Boolean network [9], [11] and Bayesian and Dynamic networks [12] and [13]. It is difficult to use probabilistic Bayesian networks to learn parameters from real data due to the huge search space of the parameters.

Authors in [14] and [15] modeled cancer as a ”stuck-at” fault in the Boolean network. Faulty logic gates represented the faulty molecules in the transduction network. Much information about cellular interaction is available in the literature; authors in [14] used this prior knowledge to generate a Boolean network from pathway knowledge using the Karnaugh map. The authors produced a Boolean network for the Mitogen-Activated Protein Kinase (MAPK) transduction pathway with this method.

In [7], and [8], authors used a combination of these Boolean networks to estimate the proportion of each subpopulation. The weights of the Boolean network,i.e., the subpopulations, were evaluated by Markov Chain Monte Carlo (MCMC) methods. The model in [7] is not parallelizable, so its computation time increases with an increase in data. The model in [8] is parallelizable, and its computation time does not increase with an increase in data as long as hardware resources are not exhausted. Still, this model on a GPU is slow if the data set size is very small. Real-world experimental data is very small, so running the parallelizable on a GPU is not the best solution. Models running on a CPU are faster than a GPU with a small dataset. In this paper, we shall show a way to run a parallelizable model on a multithreaded CPU to overcome the problem faced by a GPU with a small dataset.

M. Shell was with the Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, 30332 USA e-mail: (see http://www.michaelshell.org/contact.html).

J. Doe and J. Doe are with Anonymous University.

Manuscript received April 19, 2005; revised August 26, 2015.

III. METHOD

The goal is to estimate the proportion of the subpopulations from gene expression data. The Boolean networks are used to model each subpopulation, and the weight exerted on each subpopulation on the observables is estimated. A reasonable way to model gene expression is modelling it with a Normal distribution [7]. In [7] and [8], the probability distribution of the gene expression ration depends on the weight of each Boolean network α_i, a coefficient of variation c and expression profile d_i. d_i is a vector of length N, where N is the number of subpopulations, i.e., the number of Boolean networks. Here, i represents the i^{th} gene. The expression profile represents the transcription of the observed gene. A value of 1 in the expression profile vector represents an upregulated gene and a value of 0 represents a downregulated gene for a given stimuli.

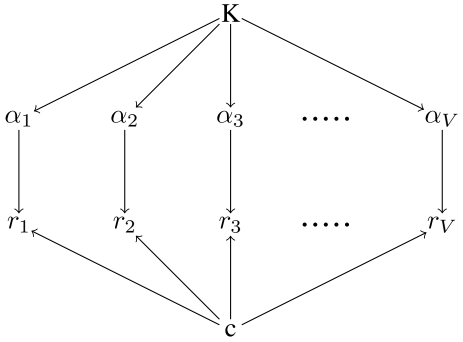

Fig. 1: The Bayesian network used in [8].

A. GPU compatible model on a multithreaded CPU

Figure 1 shows the Bayesian network of our probability model. In this model, there are V genes ,i.e, i ranges from 1 to V. Each weight vector α_i is associated with with one gene expression reading r_i for the i^{th} gene. Here, α_is depend on the vector K, and the gene expression measurement r_i depends on α_i and the coefficient of variation c. Each α_i is a vector of N elements, here, N represents the number of subpopulation or the number of Boolean networks. All the elements of α_i should sum up to one.

$$\sum_{q=1}^{N} \alpha_{i,q} = 1 \tag{1}$$

Since, r_i ranges from r_1 to r_V, so α_i ranges from α_1 to α_V.

The probability distribution of the normalized gene expression ratio r_i for the i^{th} gene from [8] is

$$\begin{aligned} P(r_i/\alpha_i, d_i, c) = \frac{m_i(r_i+m_i)}{\sqrt{2\pi}c(r_i^2+m_i^2)^{\frac{3}{2}}} \\ \times exp\left(-\frac{1}{2c^2}\frac{(r_i-m_i)^2}{r_i^2+m_i^2}\right). \end{aligned} \tag{2}$$

Here, $m_i = d_i^T \alpha_i$, c is the coefficient of variation, r_i is the gene expression ratio. The α_i for the i^{th} gene is

$$P(\alpha_i/K) = \frac{\prod_{q=1}^{N} \alpha_{i,q}^{K_q-1}}{Beta(K)} \tag{3}$$

where $Beta(K)$ is

$$Beta(K) = \frac{\prod_{q=1}^{N} \Gamma(K_q)}{\Gamma\left(\sum_{q=1}^{N} K_q\right)}. \tag{4}$$

Here, Γ is a Gamma function. K and c are the unknown parameters which is estimated by using a MCMC (Markov Chain Monte Carlo) algorithm called M-H (Metropolis-Hastings). M-H is an algorithm to sample from an unknown distribution [16], here the posterior distribution of k and α_i is unknown, so the M-H algorithm is used.

To calculate the posterior of the unknown ,i.e., K and c, prior distributions are set. The prior over $\frac{1}{c^2}$ is a Gamma distribution

$$\frac{1}{c^2} \sim \Gamma\left(\frac{\nu_0}{2}, \frac{\nu_0 c_0^2}{2}\right). \tag{5}$$

Here, Γ is a Gamma distribution. The prior over K is an exponential distribution. The means of this distribution are well separated.

The full conditional of α_i is

$$\begin{aligned} P(\alpha_i/K, c, r, d) \propto P(r_i/\alpha_i, d_i, c)P(\alpha_i/K) \\ \propto \left(\frac{m_i(r_i+m_i)}{(r_i^2+m_i^2)^{\frac{3}{2}}} exp\left(-\frac{1}{2c^2}\frac{(r_i-m_i)^2}{r_i^2+m_i^2}\right)\right) \\ \times \prod_{q=1}^{N} \alpha_{i,q}^{K_q-1}. \end{aligned} \tag{6}$$

Considering $P(K)$ as the prior distribution over K.

$$P(K/\alpha, c, r, d) \propto P(K)P(\alpha_i/K)$$

$$\propto P(K) \times \frac{1}{Beta(K)^V} \prod_{q=1}^{N} \left(\prod_{i=1}^{V} \alpha_{i,q}\right)^{K_q-1}. \tag{7}$$

The full conditional of c is:

$$\frac{1}{c^2} \sim \Gamma\left(\frac{(\nu_0+V)}{2}, \frac{(\nu_0 c_0^2 + \sum_i \frac{(r_i-m_i)^2}{r_i^2+m_i^2})}{2}\right). \tag{8}$$

Since the distribution of K and α_i are unknown so, Random Walk Metropolis Hastings will be used to sample from these distributions. Figure 2 explains the sampling from the full conditional of α_i and K on a multithreaded CPU.

Considering each α_i has three subpopulation,i.e., $N = 3$. Letting V be the number of genes. The proposal distribution

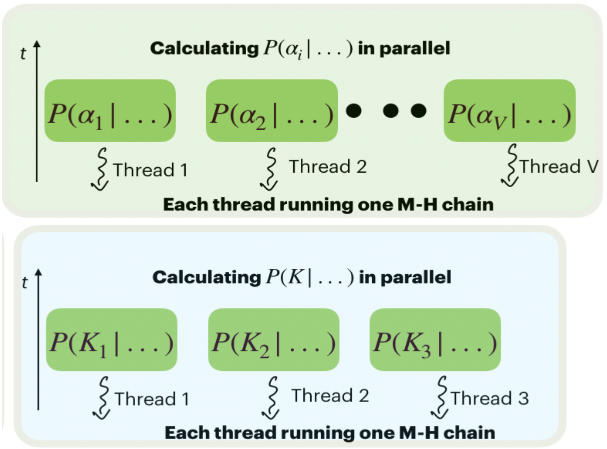

Fig. 2: Sampling of equation 6 and 7 on a multithreaded CPU.

for 6 is a Dirichlet distribution with parameter $\frac{\alpha_i}{U\alpha_i}$. The acceptance ratio is

$$R_{\alpha_i} = \frac{P(\alpha_i^*/K,c,r,\alpha_i,d)D(\alpha_i/\frac{\alpha_i^*}{U_{\alpha_i}})}{P(\alpha_i/K,c,r,\alpha_i,d)D(\alpha_i^*/\frac{\alpha_i}{U_{\alpha_i}})} \tag{9}$$

M-H algorithm is used to accept new proposals α_i^*.

The proposals K_q^* are sampled from a uniform distribution. The acceptance ratio is

$$R_K = \frac{P(K^*/\alpha,c,r,d)}{P(K/\alpha,c,r,d)} \tag{10}$$

B. Experiments with synthetic data

The algorithm was written in OpenMP and run until convergence was achieved. The synthetic data was generated by considering

$$\begin{aligned} \alpha_i &\sim D(10,6,3) \\ d_i &\sim Uniform(0,1) \\ r_i &\sim \mathcal{N}(d_i^T\alpha_i, c * d_i^T\alpha_i) \end{aligned} \tag{11}$$

Here, α_i is Dirichlet distributed with parameter K, which has been fixed to $(10\ 6\ 3)$. r_i is Normally distributed and d_i is uniform distributed.

The unknown parameters are sampled on a multithreaded CPU as shown in figure 2. Metropolis-Hastings algorithm was used to sample from the unknown distributions. The Markov chain was reached stationary at 3000 iterations, but it was run for 10000 iterations to be sure. The estimates of K came out to be $(10.6\ \ 6.1\ \ 2.8)^T$, these results are very close to the actual value $(10\ \ 6\ \ 3)^T$. The α_is are sampled from a Dirichlet distribution by considering the Ks as the parameters of the distribution, the modes of the α_i are $(0.5145\ \ 0.3743\ \ 0.1112)^T$ which is very close to the actual values $(0.5165\ 0.3746\ 0.1089)^T$.

Figure 3 shows the CPU and GPU runtime. The figure shows that for multithreaded CPU code is faster than the single threaded CPU code and the GPU code for data size less than 300.

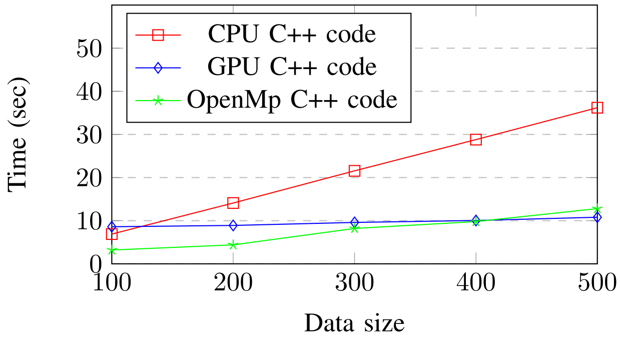

Fig. 3: Comparision of CPU and GPU Run Time for 10,000 Monte Carlo iterations

IV. Experiments with real data

To check the correctness of the model, the algorithm was run with real world experimental data collected from [7]. Samples are drawn from the posterior distribution of K using the M-H algorithm; the previous section discussed the sampling procedure. The marginal distribution of the three components of α are estimated as described before. The modes of the distribution is $(0.6161\ 0.3236\ 0.0603)^T$, which is very close to the results obtained in [7], i.e., $(0.64530\ .22550\ .1292)^T$. The faultless network has the maximum influence on the observables.

V. Conclusion

This paper addresses an important problem of designing an algorithm that can be parallelized to study cancer tissue heterogeneity. This algorithm uses prior pathway knowledge to estimate the proportion of each subpopulation. The gene expression was modeled as ratios of normally distributed random variables whose means are affected by the networks included. We also demonstrated how to use M-H MCMC algorithm on a multithreaded CPU to estimate the unknown parameters. This estimate is useful to find out the dominant subpopulation among all the subpopulations. We used the algorithm in [18] to sample from a Gamma distribution on a GPU. This helped us to parallelize the algorithm and reduce the computation time.

References

[1] P. C. Nowel, “The clonal evolution of tumor cell populations,” Science, vol. 194, no. 4260, pp. 23–28, 1976.

[2] R. A. Weinberg, The Biology of Cancer, 1st ed. Princeton, NJ, USA: Garland Science,, 2006.

[3] T. Reya, S. J. Morrison, M. F. Clarke, and I. L. Weissman, “Stem cells, cancer, and cancer stem cells,” Nature, vol. 414, no. 6859, pp. 105–111, 2001

[4] F. Barabe, J. A. Kennady, K. J. Hope, and J. E. Dick, “Modeling the initiation and progression of human acute leukemia in mice,” Science, vol. 316, no. 5824, pp. 600–604, 2007.

[5] J. C. Wang and J. E. Dick, “Cancer stem cells: lessons from leukemia,” Trends Cell. Biol., vol. 15, no. 9, pp. 494–501, 2005.

[6] L. L. Campbell and K. Polyak, “Breast tumor heterogeneity:cancer stem cells or clonal evolution?” Cell Cycle, vol. 6, no. 19, pp. 2332 – 2338, 2007.

[7] A. K. Mohanty, A. Datta, and V. Venkatraj, “A model for cancer tissue heterogeneity,” IEEE t. Bio-Med. eng., vol. 61, no. 3, pp. 966 – 974, 2014.

[8] A. K. Mohanty, A. Datta and V. Venkatraj, ”A Conjugate Exponential Model for Cancer Tissue Heterogeneity,” in IEEE Journal of Biomedical and Health Informatics, vol. 20, no. 2, pp. 699-709, March 2016.

[9] A. Datta and E. Dougherty, Introduction to Genomic Signal Processing With Control. New York, NY, USA: CRC Press, 2007.

[10] J. M. Bower and H. Bolouri, Computational Modeling of Genetic and Biochemical Networks, 1st ed. Boston, MA, USA: MIT Press, 2001.

[11] I. Shmulevich, E. R. Dougherty, S. Kim, and W Zhang, “Probabilistic Boolean networks: A rule-based uncertainty model for gene regulatory networks,” Bioinformatics, vol. 18, no. 2, pp. 261–274, 2002.

[12] N. Friedman, M. Linial, I. Nachman, and D. Pe’er, “Using Bayesian Networks to analyze expression data,” J. Comput. Biol., vol. 7, no. 3–4, pp. 601–620, 2000.

[13] M. Zou and S. D. Conzen, “A new Dynamic Bayesian Network (DBN) approach for identifying gene regulatory networks from time course microarray data,” Bioinformatics, vol. 21, no. 1, pp. 71–79, 2005.

[14] R. K. Layek, A. Datta, and E. R. Dougherty, “From biological pathways to regulatory networks,” Mol. BioSyst., vol. 7, pp. 843–851, 2011.

[15] R. K. Layek, A. Datta, M. Bittner, and E. R. Dougherty, “Cancer therapy design based on pathway logic,” Bioinformatics, vol. 27, no. 4, pp. 548–555, 2011.

[16] P. D. Hoff, A First Course in Bayesian Statistical Methods. New York, NY, USA: Springer Texts in Statistics, 2009.

[17] Frigyik, Bela A., Amol Kapila, and Maya R. Gupta. ”Introduction to the Dirichlet distribution and related processes.” Department of Electrical Engineering, University of Washignton, UWEETR-2010-0006 0006, pp. 1-27, 2010.

[18] Marsaglia, George, and Wai Wan Tsang. ”A simple method for generating gamma variables.” ACM Transactions on Mathematical Software (TOMS), vol 26, no. 3,pp. 363-372, 2000.

[19] Y. Chen, E. R. Dougherty, and M. L. Bittner, “Ratio-based decisions and the quantitative analysis of cDNA microarray images,” J. Biomed. Opt., vol. 2, no. 4, pp. 364–374, 1997

[20] Scott, David W., and Stephan R. Sain. ”Multidimensional density estimation.” Handbook of statistics, vol. 24, pp. 229-261, 2005.