2022 International Conference on Electrical, Computer and Energy Technologies (ICECET) | 978-1-6654-7087-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICECET55527.2022.9872811

Survival analysis for the identified cancer gene subtype from the co-clustering algorithm

Logenthiran Machap
Department of Computing and Information Technology
Tunku Abdul Rahman University College Johor
Segamat, Malaysia
logenthiran@tarc.edu.my

Kohbalan Moorthy
College of Computing and Applied Science
Universiti Malaysia Pahang College Johor
Pahang, Malaysia
kohbalan@ump.edu.my

Abstract— **Cancer gene subtype information is significant for understanding tumour heterogeneity. The early detection of cancer and subsequent treatment can be lifesaving. However, it is hard clinically and computationally to detect cancer and its subtypes in their early stages. Therefore, we extend the analysis and results from Machap et al. (2019), to include the Kaplan-Meier survival analysis with the integration of gene expression and clinical features data. There are two cancer datasets used for the analysis: breast cancer and glioblastoma multiforme. The luminal type was the common subtype of breast cancer, showing a higher survival rate. Whereas the Proneural subtype in glioblastoma multiforme has a little longer survival rate than the other three subtypes. These molecular differences between subtypes have been shown to correlate very well with clinical features and survival parameters to help understand the disease and develop better therapeutic targets.**

Keywords— cancer, subtype, Kaplan-Meier, gene expression, clinical features

I. Introduction

The cancer gene subtype is a type of cancer divided into smaller groups. The division is based on specific characteristics of the cancer cells. The characteristics are the physical condition of a cell under a microscope and biochemical changes in the cells for instance changes in DNA [1]. A subtype of cancer is essential to design diagnosis and decide prognosis (NCI Dictionaries). Therefore, a gene expression dataset is used to identify the cancer gene subtype. DNA sequence has unique patterns with different features called, exons, introns, and promoters. They are contributed to identifying cancer gene subtypes. Researchers are continuously searching for reliable cancer gene subtype identification methods. There are two types of approaches available for subtype identification; experimental [2] and computational approaches [3].

Over the decades, scientists have introduced in silico methods [4]. These approaches are to assist biological analysis for instance with biological network approaches. The understanding of a cell's disease processes is useful for predictive systems [2]. The computational approaches integrate gene expression with biological information. Thus it can improve the identification of disease-associated genes. Furthermore, computational approaches are more time-consuming and reduce complexity [5].

Traditional clustering methods usually do not utilize the full benefit of the gene expression data [6]. The clustering is done for genes first and then for samples or vice versa. This means a gene does not present in the other cluster. Thus, these clusters are considered low-quality clusters to be used in cancer diagnosis and prognosis. A gene can involve in more than one biological process that can be grouped into multiple clusters. In different conditions of the biological process, there are possibilities for a gene to be clustered in numerous clusters [7]. Thus, in our recent paper, [8] published a co-clustering algorithm to address this problem. Besides, further analysis is required from the subtypes produced from the improved co-clustering algorithm. Therefore, this paper is focusing on survival analysis for breast cancer and glioblastoma multiforme using some clinical features.

II. Materials and methods

Two cancer microarray datasets were used in this research. They are Breast Cancer (BRCA) and Glioblastoma Multiforme (GBM) which are publicly available at The Cancer Genome Atlas (TCGA). BRCA is obtained from [9] meanwhile GBM is from [10]. BRCA consists of 547 samples while GBM consists of 202 samples.

[11] first proposed the NCIS method to detect cancer subtypes through semi-nonnegative matrix factorization. Thus, an improved Network assisted Co-clustering for the Identification of Cancer Subtypes (iNCIS) from the existing cancer subtype identification approach is proposed. The iNCIS is proposed to generate co-clusters of cancer subtypes from cancer microarray datasets for better specificity, sensitivity, clustering quality (Rand Index and F1-measure), and classification accuracy [11]. The algorithm begins with train weights for the genes. Then, the initialization of clusters using k-means begins. This is the beginning of the co-clustering algorithm, which has two parts; the first is the objective function, and the second is optimization. The modification of iNCIS is done on the objective function that is adapted from [12]. The detected co-clusters are consistent with important genetic pathways and gene ontology annotations associated with cancer disease. The final output produced from this iNCIS primarily

Authorized licensed use limited to: Universiti Malaysia Pahang Al Sultan Abdullah (UMPSA).. Downloaded on October 31,2023 at 08:04:49 UTC from IEEE Xplore. Restrictions apply.