Decision Science Letters 12 (2023) 89–96

Contents lists available at GrowingScience

Decision Science Letters

homepage: www.GrowingScience.com/dsl

HC-UAP: Outliers detection method based-on hierarchical clustering for universally aligned time-series RNA-Seq profiles

Abedalrhman Alkhateeb[a*]

[a]*Software Engineering Department, Princess Sumaya University for Technology, Al-Jubaiha, P.o.Box 1438, Amman 11941, Jordan*

C H R O N I C L E

Article history:
Received August 22, 2022
Received in revised format:
September 30, 2022
Accepted October 13, 2022
Available online
October 13, 2022

Keywords:
Next-generation sequencing
Transcriptome
RNA-Seq
Time-series
Clustering

A B S T R A C T

Tracking abundant gene transcripts quantification over continuous cancer progression stages may reveal the mechanism of disease advancement. In this work, we profile the transcript quantification over the stages using a time-series approach, in which the stages/sub-stages of the disease are the time points, and the quantification measurements are the values. The values over time points are used to interpolate the growth of the progression using the cubic spline function. Then, the transcripts profiles are universally aligned and clustered using the time-series profile hierarchical clustering method based on the area between each pair of the aligned profiles; the method is named (HC-UAP). We compare the proposed method with a hierarchical clustering method based on Euclidean distance (HC-ED). Both methods were applied on two next-generation sequencing (NGS) prostate cancer datasets, the first from the Chinese and the second from the North American population. HC-ED clusters the dataset to find patterns while HC-UAP was able to single out outliers that trend differently in both datasets. While finding patterns in gene expression that trend over stages is the standard approach for analyzing time-series models, identifying outlier transcripts that grow differently than other transcripts can provide more details about the contribution of the mRNA transcriptional activity to the disease. They also can be a potential biomarker for the disease progression.

1. Introduction

Time-series models have been receiving attention lately in health diagnosis and prediction field (Zhang et al., 2019; Ernst et al., 2006). Researchers usually focus on clustering gene expression time series profiles (Ernst et al., 2006; Chiu et al., 2015). Rueda et al. (2007) proposed a clustering method is based on a profile-alignment approach, that minimizes the (square) area between two aligned profiles, to hierarchically cluster microarray time series data (Subhani et al., 2010). Subhani et al. (2010) introduced unsupervised machine learning model by clustering multiple aligned gene expression profiles. The method generalizes pairwise alignment to all profiles by selecting one profile as the axis to the pairwise alignments with the rest. They combined *k*-means and expectation maximization (EM) clustering with multiple alignments to cluster gene profiles time-series data (Subhani et al., 2010). Distance function is required by any time-series clustering method to find the dissimilarity between samples as well as between clusters (Davies & Bouldin, 1979). Based on the nature of the biological clustering model, it may require a specific distance function that suits best for the model (Jaskowiak et al., 2014). Distance functions are categorized into metric and non-metric methods, Euclidean distance is an example of the metric distance function. Euclidean distance has been used in gene expression clustering (Vedell et al., 2013, Ferrari & De Castro 2015). Vedell et al. (2013) used a hybrid adaptive tree cut for hierarchical clustering dendrogram method to obtain transcript abundance profiles based on Euclidean distance. Gene expression changes were identified by reverse

* Corresponding author.
E-mail address: a.lkhateeb@psut.edu.jo (A. Alkhateeb)

doi: 10.5267/dsl.2022.10.004

transcriptase poly-merase chain reaction in rat livers, then they were treated using three-time points 2, 7, and 21 days. The clusters show some patterns that can predict certain physiologic consequences of agonist treatment. Ferrari et al. (2015) proposed a clustering algorithm selection by meta-learning systems: A new distance-based problem characterization and ranking combination methods. The features are three setsof meta-attributes; the first contains the direct characterization attributes, the second applies indirect methods based on Euclidean distance, while the third combines the first and second.

Chiue et al. (2015) applied cubic B-spline interpolation on gene expression time-series data to impute the missing values. The method constructs a gene relativity graph that has sliding windows throughout gene expression profiles to cluster the time-series data. Chira et al. (2015) proposed a shape-output clustering method that studies co-expressed gene growth patterns in time-series data. The shape overtimepoints for the RNA quantification profiles have been utilized as a distance function to cluster the genes.The results show an association between gene expressions and production variables. Many researchers proposed clustering methods to capture outliers (Pamula et al., 2011; Marghny et al., 2014). Pamula et al. (2011) applied *k*-means clustering to detect outliers. Marghny et al. (2014) proposed a genetic algorithm based on *k*-means clustering to identify outliers then remove them. In this work, we are extending our previous work, which assumes that prostate cancer stage/sub-stages are the time points to model the progression of the disease (Alkhateeb et al., 2015). The assumption here is that any biological process is continuous over time. This work focuses on transcripts quantification rather than gene expression for more transcriptomic details. The outliers' transcripts behave differently throughout the stages/sub-stages are notconsidered as noise. The different trending of these transcripts may provide a biological insight for the progression of prostate cancer over the different levels of aggressiveness.

2. Materials and methods

The two NGS datasets have been downloaded from the publicly available NCBI database. The first dataset contains Eight samples from the Chinese population with NCBI SRA study number ERP000550 (Ren et al., 2012). The second dataset from the North American population with NCBI SRA study number GSE54460 contains 106 samples from the North American population (Long et al., 2014). Both works described the methods of collecting data in their geographical populations, consented to their patients, and adhered to their funding agencies rules and regulations (Ren et al., 2012; Long et al., 2014).

Table 1
The first data set tumor samples

Stage/Sub-stage	Samples
T1cN0M0	ERR031038
T2aN0M0	ERR031032
T2bN0M0	ERR031026
	ERR299297
T2cN0M0	ERR031044
	ERR299295
T3bN0M0	ERR031040
T4N0M0	ERR299298

2.1 Data pre-processing

For the first dataset, all the mRNA reads were aligned to human reference hg19 using Tophat2 (Trapnell et al., 2012), then, the aligned reads of each sample were fed into Cufflinks (Trapnell et al., 2012) to assemble the transcripts using the reads guided by the transcript annotation. Cufflinks quantifies the aligned reads on each transcript measured by Fragments Per Kilobase of transcript per Million mapped reads (FPKM). For the second dataset, the reads were aligned to reference genome hg19 and constructed the transcripts using STAR (Dobin et al., 2013) and RSEM (Li & Dewey, 2011). Fig. 1 depicts the flow of the preprocessing pipeline which has been applied on both datasets. Table 1 lists the samples from the first dataset. The sample stage/substage refers to the classification of malignant tumors (TNM), which is a global standard for classifying the extent of spread of cancer, it is widely used for solid tumors. The classification starts with T to describe the size of the tumor, and sub-stage as a lower-case letter following the number. N describes the nearby (regional) lymph nodes that are involved, but in this study, no lymph is involved, so all samples have ”0” that is following N. The last character M indicates the distant metastasis, all samples in the first dataset have not metastasized to other organs, therefore, all samples have M0. The relative abundance for each transcript in each sequenced sample is used as a measurement value of the transcript expression. Then we found the differently expressed transcripts for all the stage/sub-stages using Cuffdiff method which is part of Cufflinks package. Cuffdiff uses different statistical tests to determine whether or not a value is differentially expressed among the same transcripts values for the different stage/sub-stage. Then, any transcript that differently expressed at one or more stage/sub-stage, is considered to create a profile for it, while the remaining, which have no differently expression at any stage/substage is omitted. The total number of considered transcripts is 19,698 transcript profiles. The profile of a transcript consists of the abundances that are calculated by Cuffdiff for the transcripts at each stage/sub-stage, which is the log of the average FPKM values from all patients' samples at the same stage/sub-stage.

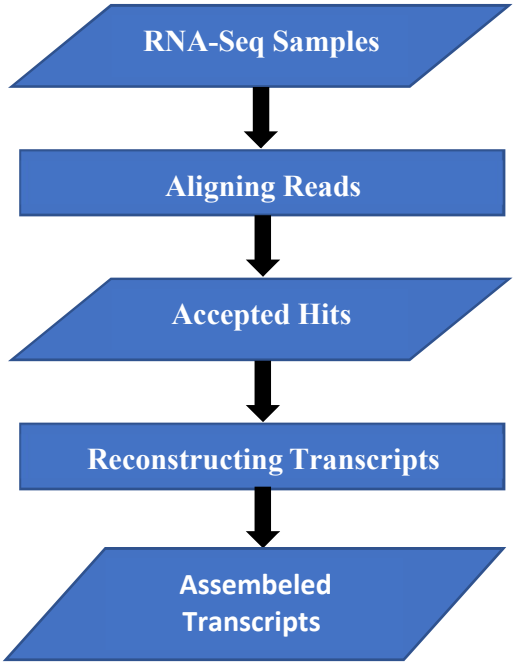


Fig. 1. The pipeline for preprocessing each sample in the data set

2.2 Natural cubic spline interpolation

Subhani et al. (2010) introduced a time-series profiles modeling using natural spline interpolations. The model interpolates the profiles by utilizing an arbitrary integral function that runs continuously ona finite interval. In our proposed method, natural cubic spline interpolates the time-series profiles, which are the quantification of transcripts on different prostate cancer stages/sub-stages. The interpolated time-series profile $x(t)$ is presented as a vector of time points$[t_1, t_2, ..., t_n]$ as:

$$x(t) = \begin{cases} x_1(t) & \text{if} \quad t_1 < t < t_2 \\ x_j(t) & \text{if} \quad t_{j+1} < t_j < t_{j+1} \\ x_{n-1}(t) & \text{if} \quad t_{n-1} < t < t_n. \end{cases} \tag{1}$$

where

$$x_j(t) = x_{j3}(t - t_j)^3 + x_{j2}(t - t_j)^2 + x_{j1}(t - t_j)^1 + x_{j0}(t - t_j) \tag{2}$$

where $x_j(t)$ interpolates $x(t)$ in interval $[t_j, t_{j+1}]$, with spline coefficients $x_{jk} \in R$, for $1 \leq j \leq n-1$ and $0 \leq k \leq 3$. At each interval $x_j(t)$, the first and second derivatives of the interpolated $x(t)$ spline equals zero, which is known as the natural condition of the spline. The interpolation of all 19,698 createdtranscript profiles are shown in Fig. 2-a).

2.3 Universal alignment

For the data set $X = x_1(t), x_2(t), ...x_m(t)$, where m is the number of time-series profiles. We universally aligned the cubic spline interpolated profiles using pairwise alignment between each profile into one specific profile $z(t)$. The alignment process shifts the profile $x(t)$ vertically towards $z(t)$ until reaching the minimal distance. The minimum distance is defined as the minimum squared error between the pair. The result of universally aligned all profiles according to $z(t)$ is seen in Fig. 2-b).

2.4 Distance function

The aim of pairwise alignment between profiles $x(t)$ and $y(t)$ is to minimize the distance between the2 interpolated curves. The vertical shifting is shifting the curves towards each other until we obtain theminimum distance a, which is the minimum square error. The method stops at the minimum distancebetween the 2 curves.

$$a_{min} = \int_0^{t_n} [x(t) - y(t)]dt \tag{3}$$

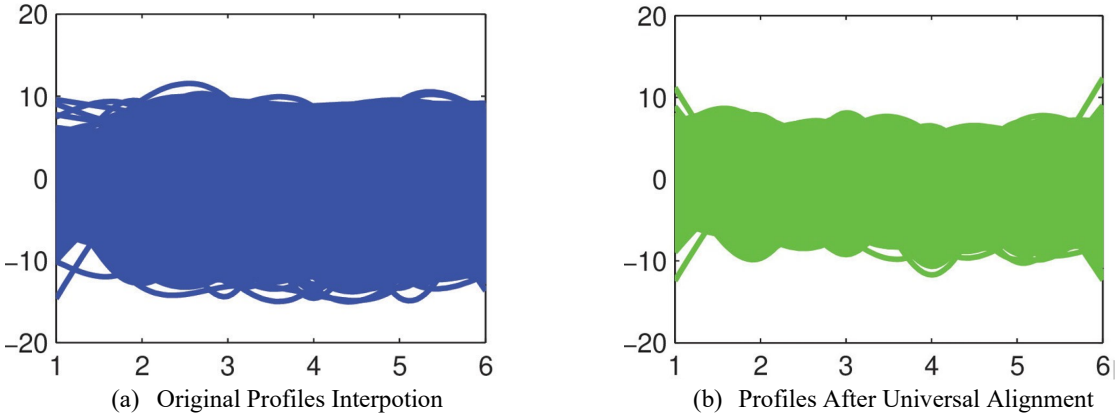
Next, the method aligns each profile to the global profile $z(t)$ to minimize the distance between them without shifting $z(t)$. Then the method calculates the distance between each pair of the profiles into thedistance matrix D as in Eq. (3).

Profile clustering

In time-series analysis, the outlier profile is defined as the one that trends differently than the rest of the profiles. The idea of detecting outlier profiles may reveal different biological functions of the other transcript profiles throughout the disease stages. In this work, outlier profiles are singled out in separate profiles, singleton clusters, or a cluster with a few profiles. In contrast, most of the profiles will remain in the background cluster, which will hold the vast majority of the profiles in the background cluster. Agglomerative clustering with maximum linkage is implemented to separate the profile with the furthest distance in a separate cluster. The method creates *m* clusters, where each profile is located in a singleton cluster, then starts merging the profiles with minimum distance in one cluster. After that,one profile may join the cluster with the minimum distance from the cluster's furthest profile. The method loops until reaching the number of desired clusters *k*, which is determined in the following section. Once the loop terminates, the outliers that trend differently from the other clusters are isolated into outlier clusters, , which are the non-background clusters..

2.5 Determining the desired number of clusters

The desired number of cluster *k* is determined using Profile Alignment and Agglomerative Clustering (PAAC) index. PAAC is proposed by Rueda et al. (2007) based on the modified version of the *I*-index cluster validity function (Maulik and Bandyopadhyay, 2002) to reduce the bias towards the smaller *k*s (Rueda and Bari, 2007) where *q* is the coefficient of normalizing the number of clusters increment and *p* the coefficient of the degree of the index. By visualizing the clusters' profiles for each *k*, and also by plotting the values of PAAC for different *k*s, we were able to select the best *k*. With coefficients values $p = 2$ and $q = 0.7$ to run PAAC index on different number of *k*, PAAC indexwas utilized on both methods on both datasets. Where PAACs for HC-UAP that is shown in Fig. 3 peak at $k = 31$ for the first dataset. PAAC values peak at $k = 8$ for HC-ED on the same dataset as seenin Fig. 4. For the second dataset, HC-UAP peaks at $k = 20$ as seen in Fig. 4 and Fig. 5, and HC-ED peaks at $k = 22$ as seen in Fig. 6.



(a) Original Profiles Interpotion (b) Profiles After Universal Alignment

Fig. 2. The result of calculating the log(FPKM) for the quantified values of the transcriptat each prostate cancer stage/sub-stage, then interpolates the 19,698 transcript profiles. (a) allprofiles before universal alignment. (b) all profiles after universal alignment

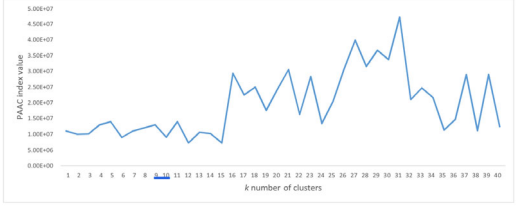


Fig. 3. PAAC values for applying HC-UAP method o the first data set

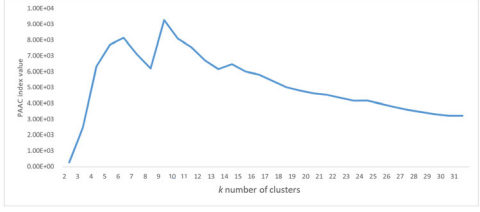


Fig. 4. PAAC values for applying HC-ED method on the first data set

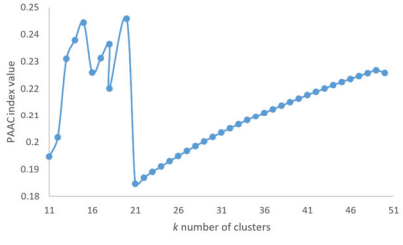


Fig. 5. PAAC values for applying HC-UAP method o the second data set

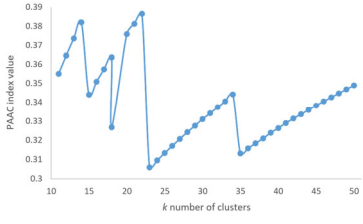


Fig. 6. PAAC values for applying HC-ED method on the second data set

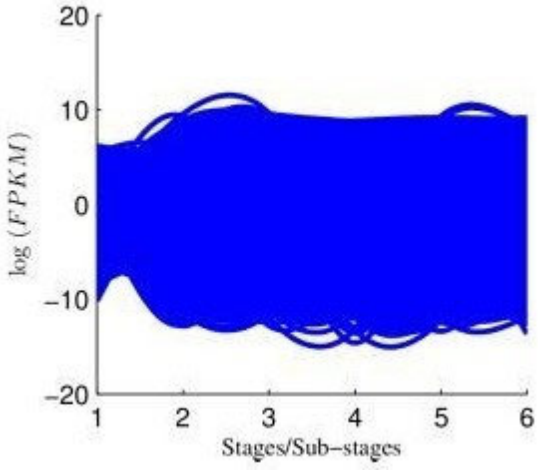


Fig. 7. The first dataset: cluster 2 - main cluster for the proposed method

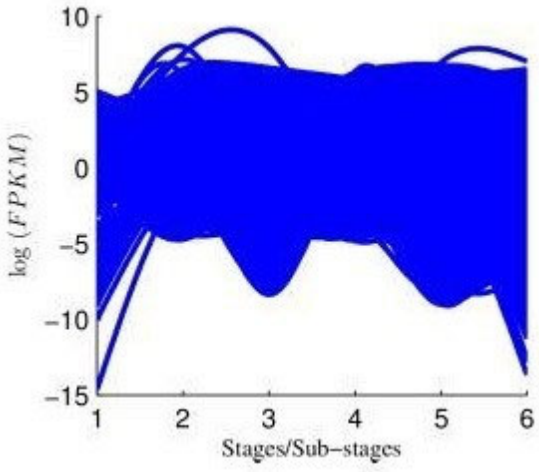


Fig. 8. The first dataset: cluster 2 - main cluster for the hierarchical clustering based onEuclidean distance method

Table 2
The outlier transcripts with their corresponding gene that are related to prostatecancer from the first dataset (the Chinese population dataset)

Gene	Transcripts	Cluster number
STMN1	NM_005563	29
CAMK2G	NM_001222	19
RUNX3	NM_004350	22
MSMB	NM_002443	31
PLA2G2A	NM_001161728	8

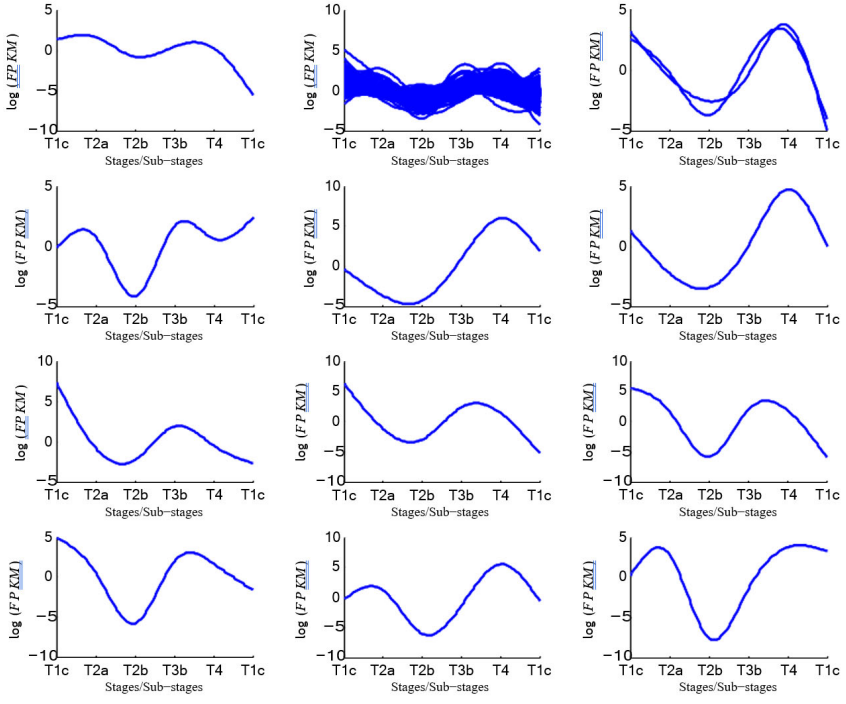


Fig. 9. The detected outliers' clusters 1-12 from the second dataset

PAAC index was utilized to carefully select the desired number of clusters k. Using validity indices for a different number of clusters, PAAC suggested $k = 31$ for the first dataset, where Cluster 2 is the background cluster that includes the majority

of the transcripts (19,656 profiles) as shown in Fig. 7. The rest are the outlier clusters that contain the outliers profiles as shown in (Alkhateeb et al., 2015). HC-ED clustered the profiles into 8 clusters based on the suggestion of PAAC analysis, which was addressed earlier. Despite having fewer number of clusters than the proposed method, the main cluster of the hierarchical method has less number of profiles (18,784) compared to the proposed method, and the main cluster of hierarchical missing many of the transcripts on the main trend as shown in Fig. 8, while the proposed method's main cluster has trimmed-looking profiles, where the outlier transcripts are removed. The second dataset has 24,368 transcripts, PAAC with the same setting as for the first dataset suggested $k = 20$. 24,348 transcripts were identified in the background cluster, the rest that are mostly singleton clusters were identified in the other 19 clusters that are seen in Fig. 9 and Fig. 10.

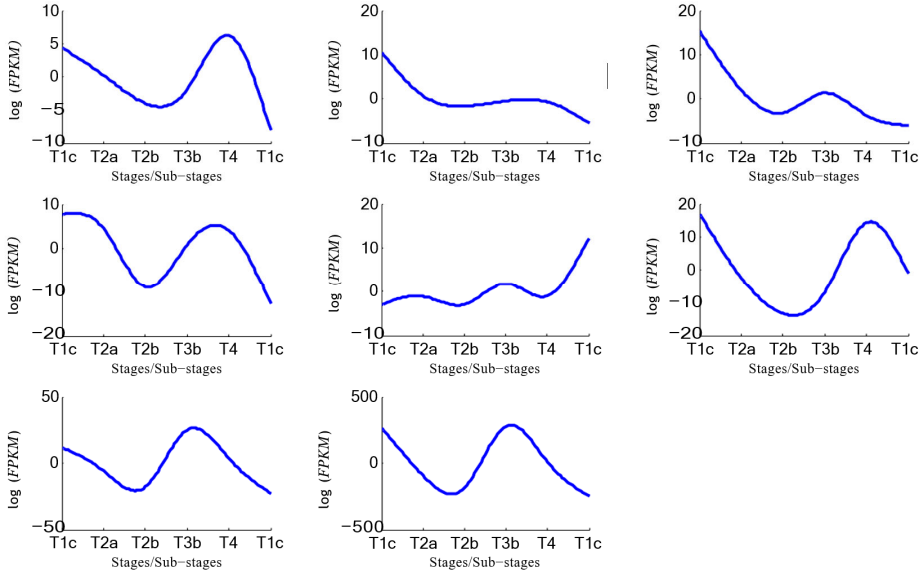


Fig. 10. The detected outliers clusters 12-20 from the second dataset

2.6 Discussion

Since the datasets have originated from different populations who have variant genomic and environmental conditions, the outlier transcripts from both datasets are different. However, both sets of outliers have shown relevancy to prostate cancer, and some of the outlier transcripts are related to another type of cancer as well. The two resulting sets of outliers are shown in Tables 2, 3 with their corresponding gene, and the cluster number in which the transcript was selected. Fig. 9 and Fig. 10 plots the outlier transcripts in their clusters. The STMN1 is a protein coding gene for Stathmin 1, an intracellular phosphoprotein that is up-regulated in prostate cancer. Increased expression levels are correlated with poor prognosis and disease progression. Microtubule assembly requires STMN1 during mitosis and constitutive phosphorylation may lead to oncogenesis. Decreased levels of STMN1 cause epithelial-mesenchymal transition and metastasis through p38 and TGF-B mechanisms. Consequently, it has been reported that the expression levels of STMN1 may be stage-dependent (Williams et al., 2012). The gene product of calcium/calmodulin-dependent protein kinase II gamma (CAMK2G) is one of four subunits belonging to the multi-functional serine/threonine-protein kinase family. CAMK2G has been reported as an enhancer of cell growth and survival in many cancers. including lung cancer, leukemia, and liver cancer (Gu et al., 2012; Meng et al., 2013; Chai et al., 2015).

Runt-Related Transcription Factor 3 (RUNX3) gene is frequently down-regulated in prostate cancer (Chen et al., 2014). Decreased levels of RUNX3 increases vascular endothelial growth factor (VEGF) secretion and thereby increasing angiogenesis. RUNX3 down-regulation also plays a role in both tumorigenesis and metastasis through dysregulation of TIMP-2/MMP-2 levels (Chen et al., 2014). The microseminoprotein beta (MSMB) gene encodes prostate secretory protein 94, a member of the immunoglobulin binding factor family. It is synthesized by the epithelial cells of the prostate gland and secreted into the seminal plasma. The result indicated that the expression of this gene is decreased in prostate cancer (Xuan et al., 1995; Sasaki et al., 1996).

The PLA2G2A gene codes for the Phospholipase A2 group 2A extracellular enzyme that plays a role in both tumorigenesis and the inflammatory response (Oleksowicz et al. 2012). PLA2G2A has been shown to be up-regulated in prostate cancer and associated with a poor response to chemotherapy as well as an overall poor prognosis. It has been suggested that the PLA2G2A enzyme might suppress genes that are induced by interferons (Fijneman et al., 2009), and is a downstream target of the HER/HER2 pathway (Oleksowicz et al. 2012). DMKN gene is associated with prostate cancer in earlier studies (Srivastava et al., 2016, Gao et al., 2017).

Srivastava et al. (2016) reported that DMKN has mutations with higher confidence at least in 14 samples out of 65 prostate samples. Gao et al. identified DMKN within a set of genes as a potential target in prostate cancer cells (Gao et al., 2017). Sayagués et al. (2010) detailed that disruption of the FAM27L gene may play a role in the malignant transformation and/or the metastasis of collateral tumors into the liver. Wang et al. (2017) reported that the nudix hydroxylase (NUDT) family of genes may have notable roles in cancer growth and metastasis. The study determined the prognostic ability of NUDT genes in clear cell renal cell carcinoma (ccRCC). Lee et al. (2017)'s results suggest that POLR2A can influence prognosis in early-stage non-small cell lung cancer (NSCLC) patients.

3. Conclusions

In this work, we modeled the prostate cancer progression using a time-series model by considering the stage/sub-stage of cancer as a time point, then interpolate the transcript growth over time using cubic spline-based on the quantification values at the various time points. A hierarchical clustering method has been used to cluster the transcript into different clusters by a full-linkage technique to discriminate the different trending transcripts. The distance used here was the minimized area under the interpolated curves of the transcripts after universally aligning them to a global transcript. The computational model was applied to two different datasets, and it was able to extract many outlier transcripts that are strongly related to the disease progression in both of them.

Acknowledgments

This work is partially funded by a seed fund by King Abdullah I School of Graduate Studies and Scientific Research.

References

Alkhateeb, A., Rezaeian, I., Singireddy, S., & Rueda, L. (2015, November). Obtaining biomarkers in cancer progression from outliers of time-series clusters. In *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 889-896). IEEE.

Chai, S., Xu, X., Wang, Y., Zhou, Y., Zhang, C., Yang, Y., ... & Wang, K. (2015). Ca2+/calmodulin-dependent protein kinase IIγ enhances stem-like traits and tumorigenicity of lung cancer cells. *Oncotarget, 6*(18), 16069.

Chen, F., Wang, M., Bai, J., Liu, Q., Xi, Y., Li, W., & Zheng, J. (2014). Role of RUNX3 in suppressing metastasis and angiogenesis of human prostate cancer. *PLoS one, 9*(1), e86917.

Chira, C., Sedano, J., Villar, J. R., Camara, M., & Prieto, C. (2015). Shape-output gene clustering for time series microarrays. *In 10th International Conference on Soft Computing Models in Industrial and Environmental Applications (pp. 241-250).* Springer, Cham.

Chiu, T. Y., Hsu, T. C., Yen, C. C., & Wang, J. S. (2015). Interpolation based consensus clustering for gene expression time series*. BMC bioinformatics, 16*(1), 1-17.

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence, 2*, 224-227.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M.,and Gingeras, T. R. Star: ultrafast universal rna-seq aligner. *Bioinformatics*, *29*(1), 15–21, 2013

Ernst, J., & Bar-Joseph, Z. (2006). STEM: a tool for the analysis of short time series gene expression data. *BMC bioinformatics, 7*(1), 1-11.

Ferrari, D. G., & De Castro, L. N. (2015). Clustering algorithm selection by meta-learning systems: A new distance-based problem characterization and ranking combination methods. *Information Sciences, 301*, 181-194.

Fijneman, R. J., Bade, L. K., Peham, J. R., Van De Wiel, M. A., Van Hinsbergh, V. W., Meijer, G. A., ... & Cormier, R. T. (2009). Pla2g2a attenuates colon tumorigenesis in azoxymethane-treated C57BL/6 mice; expression studies reveal Pla2g2a target genes and pathways. *Analytical Cellular Pathology, 31*(5), 345-356.

Gao, W., Lam, J. W. K., Li, J. Z. H., Chen, S. Q., Tsang, R. K. Y., Chan, J. Y. W., & Wong, T. S. (2017). MicroRNA-138-5p controls sensitivity of nasopharyngeal carcinoma to radiation by targeting EIF4EBP1. *Oncology Reports, 37*(2), 913-920.

Gu, Y., Chen, T., Meng, Z., Gan, Y., Xu, X., Lou, G., ... & Xu, R. (2012). CaMKII γ, a critical regulator of CML stem/progenitor cells, is a target of the natural product berbamine. Blood, *The Journal of the American Society of Hematology, 120*(24), 4829-4839.

Jaskowiak, P. A., Campello, R. J., & Costa, I. G. (2014, January). On the selection of appropriate distances for gene expression data clustering. *In BMC bioinformatics*, *15*(2), 1-17.

Lee, J. H., Yoo, S. S., Hong, M. J., Choi, J. E., Lee, S. Y., & Park, J. Y. (2017). Association between polymorphisms in microRNA target sites and survival in early-stage non-small cell lung cancer. *Annals of Oncology, 28*, v456.

Li, B. and Dewey, C. N. Rsem: accurate transcript quantification from rna-seq data with or withouta reference genome. *BMC bioinformatics*, *12*(1), 1–16, 2011.

Long, Q., Xu, J., Osunkoya, A. O., Sannigrahi, S., Johnson, B. A., Zhou, W., ... & Moreno, C. S. (2014). Global Transcriptome Analysis of Formalin-Fixed Prostate Cancer Specimens Identifies Biomarkers of Disease Recurrence Biomarkers of Recurrence in Prostate Cancer. *Cancer research, 74*(12), 3228-3237.

Marghny, M. H., & Taloba, A. I. (2014). Outlier detection using improved genetic k-means. arXiv preprint arXiv:1402.6859.

Maulik, U., & Bandyopadhyay, S. (2002). Performance evaluation of some clustering algorithms and validity indices. *IEEE Transactions on pattern analysis and machine intelligence, 24*(12), 1650-1654.

Meng, Z., Li, T., Ma, X., Wang, X., Van Ness, C., Gan, Y., ... & Huang, W. (2013). Berbamine Inhibits the Growth of Liver Cancer Cells and Cancer-Initiating Cells by Targeting Ca2+/Calmodulin-Dependent Protein Kinase IIBerbamine Inhibits Liver Cancer through CAMKII. *Molecular cancer therapeutics, 12*(10), 2067-2077.

Oleksowicz, L., Liu, Y., Bracken, R. B., Gaitonde, K., Burke, B., Succop, P., ... & Lu, S. (2012). Secretory phospholipase A2-IIa is a target gene of the HER/HER2-elicited pathway and a potential plasma biomarker for poor prognosis of prostate cancer. *The Prostate, 72*(10), 1140-1149.

Pamula, R., Deka, J. K., & Nandi, S. (2011, February). An outlier detection method based on clustering. In 2011 second international conference on emerging applications of information technology (pp. 253-256). IEEE.

Ren, S., Peng, Z., Mao, J. H., Yu, Y., Yin, C., Gao, X., ... & Sun, Y. (2012). RNA-seq analysis of prostate cancer in the Chinese population identifies recurrent gene fusions, cancer-associated long noncoding RNAs and aberrant alternative splicings. *Cell research, 22*(5), 806-821.

Rueda, L., & Bari, A. (2007, December). Clustering temporal gene expression data with unequal time intervals. *In 2007 2nd Bio-Inspired Models of Network, Information and Computing Systems (pp. 192-199).* IEEE.

Sasaki, T., Matsumoto, N., Jinno, Y., Niikawa, N., Sakai, H., Kanetake, H., & Saito, Y. (1996). Assignment of the human β-microseminoprotein gene (MSMB) to chromosome 10q11. 2. *Cytogenetic and Genome Research, 72*(2-3), 177-178.

Sayagués, J. M., Fontanillo, C., Abad, M. D. M., Gonzalez-Gonzalez, M., Sarasquete, M. E., Chillon, M. D. C., ... & Orfao, A. (2010). Mapping of genetic abnormalities of primary tumours from metastatic CRC by high-resolution SNP arrays. *PLoS One, 5*(10), e13752.

Srivastava S. K., Dobi, A., Petrovics, G., Werner, T., Seifert, M., & Scherf, M. (2016). Prostate cancer gene profiles and methods of using the same. US Patent App. 15/108,909

Subhani, N., Rueda, L., Ngom, A., & Burden, C. J. (2010). Multiple gene expression profile alignment for microarray time-series data clustering. *Bioinformatics, 26*(18), 2281-2288.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., ... & Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols, 7*(3), 562-578.

Vedell, P. T., Lu, Y., Grubbs, C. J., Yin, Y., Jiang, H., Bland, K. I., ... & Lubet, R. (2013). Effects on gene expression in rat liver after administration of RXR agonists: UAB30, 4-methyl-UAB30, and Targretin (Bexarotene). *Molecular pharmacology, 83*(3), 698-708.

Wang, Y., Wan, F., Chang, K., Lu, X., Dai, B., & Ye, D. (2017). NUDT expression is predictive of prognosis in patients with clear cell renal cell carcinoma. *Oncology letters, 14*(5), 6121-6128.

Williams, K., Ghosh, R., Giridhar, P. V., Gu, G., Case, T., Belcher, S. M., & Kasper, S. (2012). Inhibition of Stathmin1 Accelerates the Metastatic ProcessSTMN1 Accelerates the Metastatic Process. *Cancer research, 72*(20), 5407-5417.

Xuan, J. W., Chin, J. L., Guo, Y., Chambers, A. F., Finkelman, M. A., & Clarke, M. W. (1995). Alternative splicing of PSP94 (prostatic secretory protein of 94 amino acids) mRNA in prostate tissue. *Oncogene, 11*(6), 1041-1047.

Zhang, Z., Xu, J., Tang, J., Zou, Q., & Guo, F. (2019). Diagnosis of brain diseases via multi-scale time-series model. *Frontiers in Neuroscience, 13*, 197.